

# AI-Inspired Biology: Does AI Have Something to Contribute to Biology?

Jackie Chappell and Susannah Thorpe<sup>1</sup>

## Abstract.

Biology, and particularly the study of ‘natural intelligence’, has long provided diverse forms of inspiration for AI and robotics researchers. However, instances of biologists gaining inspiration from AI have been less common. In this paper (written as an introduction to the AI-Inspired Biology Symposium), we argue that there are many ways in which biologists interested in natural intelligence can learn from AI, and outline the different kinds of contributions that AI can make. We describe some of the open, unsolved problems in animal cognition — including more detailed case studies of two problems we are particularly interested in: orangutan locomotion and New Caledonian crow tool use — and highlight the ways in which we believe AI can contribute to understanding and eventually solving these problems. Finally, we discuss the potential barriers to collaboration between the two fields, and ways in which we can make progress and collaborate to our mutual benefit.

## 1 INTRODUCTION

Imagine watching an orangutan (*Pongo abelii*) slowly making his way through the tangled canopy of a Sumatran rainforest. He arrives at a gap which contains an isolated young tree with a relatively thin, flexible trunk, then reaches out to grasp the trunk. As he pulls himself into this tree, he shifts his body weight — somewhat like a child on a swing — to oscillate the tree until its sway brings him within grasping distance of the branches on the far side of the gap.

A female elephant (*Loxodonta africana*) and her calf are bathing and drinking in a muddy waterhole with the rest of their social group. When it is time to leave, the adults can step over the muddy banks easily, but the small calf keeps slipping back into the water as she tries to climb out. Observing her offspring’s difficulty, the mother uses her powerful trunk and tusks to break down the mud bank to form a gentle gradient, which her calf then climbs with ease.

A New Caledonian crow (*Corvus moneduloides*) breaks a twig off a tree, carefully trims all leaves and side branches from the twig, then hops over to a rotting log. It selects a hole in the log and inserts the stick, jiggling it rapidly up and down. Inside the hole a Cerambycid larva mounts a defence of the burrow, trying to bite the stick with its large mandibles. Like a fisherman, the crow needs to move the stick to incite the larva to bite it, but if it pulls the stick away too violently, the larva will release its grip and drop back into the hole. When the crow detects that the larva has bitten on to the stick, it pulls both rapidly out of the hole and eats the larva.

Such examples of animals<sup>2</sup> solving problems posed by their environments are numerous, and more are being discovered as the detail with which we are able to observe animal behaviour in natural environments improves. ‘Natural intelligence’ (henceforth NI) tends to be characterised by complexity, flexibility, robustness and richness. In the examples above, the animals had to adjust their behaviour depending on the current context, plan a complex sequence of actions, and perhaps imagine the effect of their planned action on the environment. However, the extent and type of complexity involved differs. For example, the orangutan has to understand itself as a causative agent in physically altering its physical environment. In contrast, the New Caledonian crow faces somewhat less complexity, and yet it must still adjust its actions in a very sensitive way, depending on the behaviour, orientation and location of the larva. In the case of the elephant, she needed to adjust her behaviour by appreciating the difference between her own motor competences and those of her offspring. Some of these examples involve ‘online’ intelligent control of processes in the environment as they are happening. Others seem to involve ‘offline’ intelligent consideration of possible processes that are not occurring, but might occur and produce effects. In the former, the detailed morphology and fine tuning of sensory motor signals of the animal are directly involved. In the latter more abstract (possibly symbolic) information processing is required. In contrast the products of artificial intelligence research (AI) in robotics, — despite decades of research — tend to be comparatively simple and restricted, and do not cope well with novel situations, or where the parameters of the task are not known in advance [27]. This is certainly not a reflection of the quality of the field of AI or its researchers, but demonstrates the enormity of the task of replicating even a tiny subset of the capabilities of animals. After all, evolution has had a head start of millions of years.

While biologists now have quite a good understanding of the range and complexity of behaviours shown by animals, we are a long way from being able to explain *how* the information processing systems underlying such behaviour function, or even to describe the structure of such systems. In contrast, concepts in AI developed to understand complex systems might provide precisely the tools we need (or at least important steps towards such tools) to make progress in this area. In this paper, we will discuss the reasons why we are arguing that AI can inspire biology, and propose some examples from our own work. Our main focus is on the role for AI’s conceptual and analytical tools, but attempting to build systems (or parts of systems) as physical robots or simulations probably will also have a role.

<sup>1</sup> School of Biosciences, University of Birmingham, UK, email: j.m.chappell@bham.ac.uk

<sup>2</sup> Unless we distinguish explicitly between human and non-human animals, we include humans when referring to ‘animals’.

## 1.1 What problems do biologists need help with?

As biologists, there are a number of different levels at which we must consider cognitive systems in NI. At the evolutionary level, we need to understand the selection pressures promoting NI, and any evolutionary trade-offs involved [4]. Since the animal's environment<sup>3</sup> poses the problems that evolution must solve in order for the animal to reproduce, understanding the scope and nature of those problems is key to understanding a species' evolution. At the next level, we need to understand the structure of the information processing system in particular species: for example, in what ways can animals process, store and reuse information in new contexts? Finally, at the level of the individual animal, we need to understand how the system is constructed, shaped and fine-tuned by both development and learning. This includes both:

- how the genome together with epigenetic processes control the construction of the organism as an extremely complex physical machine, including self-maintaining and self-repairing mechanisms
- how the genome together with epigenetic processes control the development of increasingly complex information processing systems whose relationship to the physical machinery may be very obscure (as it is in complex computer-based systems).

At all the levels described above, the central problem with understanding NI is that cognition is a 'black box' system: we know (in most cases, but see section 2.4) what information passes in to the system, and we can record the animal's subsequent behaviour, but the information processing system is largely hidden. The parts of this system also operate in parallel and concurrently, with complex interactions between parts. Some changes to the system may only alter the behaviour or interaction between internal systems or virtual machines [24, 8], with no obvious external change. As a simple example, this can occur when an animal has learned something about the temporal relationship between a stimulus and a reinforcement event, but because the stimulus occurs *after* the event, there is no benefit in making a response. This has been demonstrated in rats, which can be shown to have learned about the temporal relationship between a light cue and an electric shock (where the shock precedes the light), even though they do not show any response until given the opportunity to respond to a second stimulus (a tone), preceding both the light and the shock [1]. The experiment shows that — even though they had never experienced the tone and shock paired together — they knew about the temporal relationship between the two.

Our understanding is also limited by the fact that our own cognitive processes differ from those of other animals, which can sometimes lead us to make unwarranted assumptions [19]. In addition, biological explanations for the evolution of cognitive abilities in animals tend to be rather specific to particular species or higher taxonomic grouping. While this practice can be justified because evolution shapes cognitive systems (and morphology) according to animals' precise ecological circumstances and phylogeny, it does not allow us an overview of how evolution has shaped cognitive systems in a range of species, or why other possible 'solutions' were not adopted by evolution. Finally biologists (and psychologists) often neglect individual variation, treating it only as a statistical inconvenience, rather than a valuable source of information about the ways in which ontogeny and learning can shape NI (see also section

<sup>3</sup> When we refer to 'environment' we include not only the physical habitat of the animal, but also members of its own species, its prey, predators and so on.

2.5 for a related problem). If one individual solves a particular problem, while the others fail, it is important to examine the strategies of each of the subjects in detail and look for relevant differences. For example, imagine you have presented subjects with the problem of removing the lid of an opaque box, which is fixed to the body of the box with a number of bolts, only some of which are functional. 'Distractor' bolts are in fact two separate bolts, aligned vertically, but not connected inside the box. The successful subject might have avoided trying to remove the distractor bolts (thus solving the problem) because it noticed that when the bottom of the bolt was rotated, the top part did not also rotate, suggesting that they were not connected. Careful observations such as differences between subjects can potentially provide important information about the way in which relevant information is collected and processed by individuals.

## 1.2 Examples of open problems in animal cognition

In addition to the broad problems outlined in section 1.1, there are a number of other more specific unsolved problems in animal cognition, all of which would — in our opinion — benefit from insights derived from the field of AI.

1. It is generally agreed that there is a large difference between the cognitive capabilities of human and non-human animals. However, it is not clear whether the difference is a mainly quantitative one (humans are better at a greater number of tasks than non-human animals) or a qualitative one (cognitive systems in humans and non-humans are of fundamentally different kinds), or a mixture. Whatever the nature of the difference turns out to be, why and how did it evolve? What kinds of selective pressures promote advanced cognition, and are they the same in all animals? Is the real difference between our cognitive abilities and those of other animals flexible, extensible, domain-general abilities, rather than domain-specific adaptations for problem solving in a particular domain? How do we distinguish between examples of domain-general intelligence and those of domain-specific intelligence?
2. How do some organisms acquire complex cognitive capabilities so quickly [26, 8, 25]? What is the structure of such a cognitive architecture [16]? The provision of 'innate modules' (requiring little or no environmental input) would provide rapid acquisition of skills though without much flexibility, but learning from a "blank slate" state (which would provide more flexibility) is both computationally and biologically implausible [18]. Is there a process which could provide the best of both worlds [8], by structuring learning in such a way that existing competences could be combined in a manner that is specifically influenced by the environment, to form qualitatively new abilities? As biologists, it seems implausible to us that evolution would have neglected to make use of the unchanging properties of the environment, which remain stable from generation to generation, and bias animals' learning with certain 'expectations' about these properties. We know from research on children that they have certain expectations about solidity and continuity of surfaces [28], and it would be surprising if some non-human animals did not share some of these expectations, along with others concerning geometry and topology of objects, connection, contact and so on. If there is any truth in Chomsky's well-known (and much disputed) claim that humans have an innate language learning mechanism, then it may turn out to be a special case of the class of mechanisms supporting "meta-configured" competences described in [8].
3. If having a big brain and complex cognition is so evolutionarily advantageous, why don't more animals have a large brain and

complex cognition? We need to investigate what selective advantages advanced cognition brings, and what the constraints on evolving such abilities are (see [4, 11] for two opposing views on this topic).

## 2 HOW CAN AI HELP?

### 2.1 The ‘designer stance’

In NI, evolution is the ‘designer’, though it is not (of course) working towards a predetermined final design. As biologists, we tend to observe the behaviour of the animal, and then think about how the animal’s environment and phylogenetic history might have selected for that behaviour. In contrast, when taking a ‘designer stance’ [12, 18, 23], we can reverse this process and attempt to determine the ‘requirements’ of the environment. In other words, we can evaluate the design problems posed by the animal’s environment, and then consider the range of designs which could fulfil these requirements, comparing our findings the actual designs revealed by observations and experiments on animals (see [18]). It can also be instructive to consider the opportunities *not* taken by evolution. For example, in an analogous exercise, Raup [20] developed a model to explore the morphological diversity of mollusc shells, varying parameters of shell geometry to produce a complete theoretical ‘morphospace’ for all possible mollusc shells. While there are living or fossil examples representing many of the regions of the morphospace, some parts are empty, raising interesting questions about why those designs were not exploited by evolution. This model considered only the range of designs possible, not the requirements imposed by the environment which determine functional shell designs, but considering requirements and designs in tandem would be instructive.

### 2.2 The value of construction

A benefit that was not obvious to us before we started to interact with AI researchers is the value of constructing a model (real or simulated) over attempted deduction of mechanisms from behaviour. As biologists, we are used to working with fully competent animals, so it is very easy to overlook subtle but important details and processes when thinking about the mechanisms involved in a particular behaviour. However, thinking about how to construct a complete or partial working system quickly reveals the gaps in one’s knowledge. For example, one has to consider the entire process, from defining what sensory information is collected and stored by the animal, through to the motor actions the animal takes on the environment (see also section 2.4), and this can reveal hitherto unnoticed gaps in understanding. We have personally found that — even as a thought experiment — it is very illuminating to think through the process of how you would implement a particular behaviour, guided and prompted by AI researchers. However, attempting to construct a physical robot or simulation is even more revealing. For example, when building a robot which interacts with the world, AI researchers have to consider how to filter the incoming information and attend preferentially to the ‘important’ information. Since we do not normally need to be concerned with attention as biologists<sup>4</sup>, we often neglect to think about how animals filter sensory information and direct their attention, nor do we generally consider what kinds of information can be considered ‘important’ to the animal.

<sup>4</sup> Psychologists do consider many aspects of attention, but usually in a laboratory, not a ‘field’ environment.

### 2.3 Generating hypotheses

Interacting closely with AI researchers gives us a way of evaluating alternative theories. Of course, the domains of NI and AI are radically different (at least at present), so results of experiments in either domain cannot be compared directly. However, we can capitalise on the complementary strengths and weaknesses of NI and AI to help generate new hypotheses for both systems. NI systems are rich and complex, but we know very little about the mechanisms involved, while AI systems are poorer and less complex, but because we have constructed the system ourselves, we know about the mechanisms involved in great detail. In addition, we can perform experiments with AI systems (such as removing parts of the mechanism, depriving the system of particular information and so forth) which are impractical or unethical to do with animals. Information about the results of such experiments, the kinds of tasks on which the AI systems succeed or fail, and the nature of such successes and failures, can all help us to generate new hypotheses about NI systems.

### 2.4 A systems approach

Designers of robots need to think about the entire set of interconnected, interacting processes involved in obtaining information from the environment, processing it, acting on it, and perceiving the resultant change in the world: they are responsible for constructing those systems and getting them to interact. Biologists, with the luxury of studying fully-formed, fully-capable animals, tend to specialise in studying one part of the system. This is changing gradually with collaborations between biologists, psychologists and physiologists, but collaborating with AI researchers can help us to think about the processes in a more abstract way, later applying that abstract knowledge to what we already know about NI systems, and helping to identify gaps in our knowledge.

### 2.5 A qualitative approach

Experiments in animal cognition tend to focus on establishing that species *A* has capability *X* but species *B* does not, while experiments in developmental psychology sometimes focus on the age at which children transition from not having to having capability *X*. We argue that while it is important to establish differences between species and developmental transitions within a species, it is equally vital to learn about the details of the ways in which that capability is built (what are the necessary conditions, how does it vary with developmental conditions), under what conditions it fails and so on. In other words, we need to develop the kinds of processes and logic used by AI researchers when debugging faulty programs.

## 3 EXAMPLES

### 3.1 Orangutan locomotion

To return to the orangutan who slowly meandered through the canopy of the Sumatran rainforest at the start of our paper: how can AI help us to understand his biology? He belongs to the largest mammalian species to live a habitually arboreal lifestyle; adult male orangutans may weigh up to 100kg, adult females about 40kg, and juveniles about 25kg. How do they move such a large body mass through the canopy safely to feed and to reproduce? How do orangutans deal with the problems of arboreal living compared to other species? Is this environment cognitively challenging and if so, what, if anything, can we learn about NI in orangutans; the selective pressures behind its

development and the evolutionary tradeoffs that have occurred and have thus allowed the continuation of the species.

The most challenging components of life in the forest canopy for a large-bodied habitually- arboreal species are the compliance of the branches and lianas they must use to support their body mass and the gaps in the canopy that they must traverse without descending to the ground, both to avoid ground dwelling predators such as the Sumatran tiger and to avoid the enormous energetic cost of repeatedly descending to the ground and opposing gravity to climb back up into the next tree [31]. Even the narrowest gaps between trees lie between very flexible terminal branches, which bend substantially under the animal's weight, and vary enormously in their properties.

Orangutans deal with the mechanics of crossing gaps and moving on compliant branches in a number of ways. Unlike all other arboreal primates and marsupials that deal with branch instability by walking on all fours on top of the branch but with very bent elbows and knees to bring their centre of mass close to the support and therefore reduce branch vibrations, orangutans employ an upright, mostly suspensory clambering type behaviour [17, 21] in which they use all their limbs to support the body at any angle [29, 32, 33]. In this technique they increase stability because they have, in effect, already fallen off the support and their centre of mass is thus directly under the branch bearing their weight. Balance and increased stability are further achieved through long contact times between multiple limbs and multiple supports and locomotor behaviours that lack regular limb sequences, which serves to avoid the risk of resonance in branch-sway caused by high-frequency, patterned gait [33]. For crossing large gaps, orangutans use tree sway, the behaviour we described earlier, where they oscillate flexible tree trunks with increasing magnitude until they can reach across gaps in the canopy to transfer into the next tree [5, 29, 30].

Thus, we now know a great deal about the biomechanics of how orangutans interact with their habitat [6, 29, 30, 30, 31, 33], but this tells us very little about how orangutans understand the affordances of this physically highly complex habitat and/or make decisions about how they will interact with it. Nor does it take into account the requirements posed by the animals' environment in its wider sense. This includes social constraints: even for a relatively solitary animal like the orangutan, access to preferred food resources is still dictated by a dominance hierarchy and the individual's social status and is therefore still likely to influence the way in which the animal can interact with its habitat. Other environmental considerations are infant care during arboreal travel (which increases both the cognitive and physical demands on the mother, and also provides complex learning opportunities for infants), the presence of ground dwelling predators, and ontogeny since both physical size and cognitive ability change dramatically through ontogeny but the animal must still be able to move safely through the forest to feed. Adopting the designer stance and other AI methodologies — both in theoretical analysis of problems and in attempting to design working robots or simulations — can suggest new experimental approaches that might be useful to test which aspects of a habitat are problematic, and to compare and evaluate the solutions that are used versus the solutions that are possible but have not been adopted.

One of the most interesting aspects of gap crossing in orangutans is that the geographical separation between the tree they are in and the tree they want to be in on the other side of the gap is usually too large for them to be able to experiment with the supports' responses to loading in the destination tree. Gap crossing decisions can therefore only be made by prior learning of how different gaps transform during locomotion or by using more abstract knowledge of support

mechanics and their affordances to predict how the gap will transform, before loading the supports. We propose that the combinatorial diversity of 3-D configurations of gaps (with varying gaps, numbers of possible supports, diameters, orientations, etc.), and the high costs of failed attempts (falls are much more likely to be fatal for animals of large body mass), rule out associative learning (e.g. that based on repeated retinal patterns), as the sole learning mechanism.

AI can offer a great deal here. The complexity of this system may mean that constructing an AI model is perhaps some way down the line, but again the designer stance could greatly help us understand the problems from the animals' perspective without falling into the trap of assuming there is only one way of perceiving the problems of gap crossing and deciding the most appropriate approach — the human way [23]. Indeed, the problems associated with human bias are even more severe in this scenario because orangutans travel 20-30m up in the forest canopy which means that few researchers can ever have even seen the problem from the same physical perspective as the animal, let alone from the same cognitive perspective.

A key question is if orangutans can contemplate the consequences of their locomotor actions and support behaviour in advance, then how far in advance are they able to do so? The distinction in the AI literature between reactive and deliberative planning [3] is relevant here. In reactive planning, the planner chooses only the most appropriate next action to get closer to achieving the goal. If orangutans follow the same geographical travel routes between major food resources, they might only need to make small adjustments for which reactive planning would be sufficient. However, the canopy of tropical rainforest is subject to frequent structural changes as trees and lianas grow, break or fall. These can result in large gaps in the canopy that arboreal animals must circumvent and in the loss of key feeding trees, thus changing routes travelled fundamentally. Since Sumatran orangutans rarely seem to encounter arboreal 'dead-ends' (Thorpe, pers. obs.), it is likely that they employ some level of deliberative planning, where they perceive a desired goal/sub-goal and construct a series of steps (before action is taken) to achieve it. Unless this sequence has been previously learned or is genetically encoded, deliberative planning implies that orangutans reason about the outcomes of alternative possible sequences of planned actions, and use these predictions to evaluate the best course of action.

But recent developments in AI enable us to take this process even further. There tends to be a natural inclination in biology and many other disciplines to try to divide natural phenomena into clear-cut, binary alternatives. The innate *versus* learned distinction is one such false dichotomy in biology, and the division between reactive and deliberative planning can be seen as another. In reality, when we consider the problem using the designer stance, it is likely that there are many different kinds of behaviour, which can be classified in a number of different ways, forming different kinds of organisational structure. There is an extended discussion of this topic in the paper by Sloman in these Proceedings.

We suggest that, rather than trying to divide up competences into inappropriate binary divisions, we should consider asking questions about how they evolved and the evolutionary selection pressures involved, what kinds of information and neural mechanisms are involved in their expression, and what kinds of architectures could support such types of competence. In addition, we need to think about variation between individuals and species, and how this relates to our taxonomy of the phenomenon. This approach is potentially a much more powerful method to understand not only the nature and extent of deliberative planning in orangutans (and other species), but also in addressing whether their solutions to arboreal living really differ as

much as their differences in morphology from other species would suggest.

### 3.2 Tool use in New Caledonian crows

New Caledonian crows (*Corvus moneduloides*) are highly unusual among birds because they make and use tools in rather sophisticated ways. Not only do they use sticks to extract larvae from rotten logs as described earlier [13], but they also make and use a variety of other tools [15, 34], including a multi-step process involving cutting and tearing of *Pandanus spp.* leaves to form stepped tools [14]. Tools are an integral part of their foraging behaviour, and are used frequently [2]. However, there is substantial variation in both the amount of time that individuals spend using tools, and their subsequent foraging success [2]: why do crows differ in their ability to use tools, and what factors might influence this during development and learning?

What, if anything, do New Caledonian crows (hereafter, ‘crows’) understand about the process of making and using tools, and the physical laws and causal regularities that underlie such behaviour? As discussed in sections 1.1 and 2.5, rather than attempting to demonstrate that crows either have or do not have a specified form of ‘understanding’ of the problem, we need to investigate details of the way in which they process information, including probing the kinds of tasks they tend to either succeed or fail on, the kinds of errors they make under which circumstances, and the kinds of representations of the problems that they might form [9].

Since crows use a variety of tools, one might assume that — like humans — they use each type of tool for a particular purpose, and that they select tools according to the kind of task they are faced with. One basic aspect of selectivity is that crows may choose tools with appropriate dimensions for the task facing them. Since they tend to probe for invertebrates and other food items in holes in logs, the length and diameter of tools might both be important for their function. We tested this by providing a tool using opportunity in captivity, in which the provided tools varied in length, and the distance of a food reward from the open end of a tube varied unpredictably [10]. This meant that on any one trial, some of the tools would match or exceed the distance to food (and thus be functional), while others would be too short. Both crows tested showed some evidence for selectivity, choosing tools which either matched the distance to food, or were the longest tool, significantly more often than would be expected by chance [10]. In another experiment, we probed selectivity during manufacture of tools by providing branches of oak in full leaf as the raw material for tools with which to push food out of a tube, where the aperture varied in diameter from trial to trial [7]. Again, both crows varied the maximum diameter of the tool they made in accordance with the diameter of the tube, making broader tools when the diameter of the aperture was larger [7]. Recent evidence from the field has shown that selectivity is not restricted to captivity. Crows do not appear to choose their tools randomly from the pool of available material, and longer tools tend to be found left in deeper holes [2].

Faced with such situations, why do crows not use or manufacture the longest/narrowest tool possible, since that strategy should always result in success? In trying to understand their choices, we face similar problems to those discussed earlier in relation to orangutan locomotion (section 3.1): we tend to view the problem from a human perspective, neglecting differences in both information processing and the ergonomics of tool use for crows. However, we can speculate about the possible disadvantages of using such a heuristic when selecting tools. Longer tools are inevitably heavier, and may be unwieldy to use precisely, particularly when one considers the length

of the tool relative to the bird’s body length, and that it is held in the beak, not the hand or foot. Similarly, narrower tools might be more difficult or time-consuming to make, and may break more easily.

We also know that crows show a form of creativity, making tools from materials with unfamiliar properties and using different manufacturing processes than they ordinarily use. One crow spontaneously bent a piece of straight wire into a hook to retrieve a bucket containing food from a vertical tube [35], despite the lack of similar pliable materials in the wild. Furthermore, she was able to *unbend* metal strips to retrieve food, showing a number of different techniques to achieve this [36]. Nevertheless, such experiments can sometimes reveal conflicting evidence for ‘understanding’, reinforcing the point that we need to study the details of the precise ways in which individuals attempt to solve such problems, in what circumstances this process fails, and how they gather information about possible solutions [36]. Only then will we be in a position to specify requirements for satisfactory AI models of their competences.

Returning to the vignette of New Caledonian crow tool use we presented in section 1, we have recently investigated the detail of crows’ use of tools while fishing for larvae (Trosianko, Rutz, Bluff, Kacelnik and Chappell in prep.). Since the larva moves independently, the crow must alter its probing behaviour with the tool depending upon the precise behaviour of the larva. Furthermore, the constraints of the task are such that it must combine restricted visual information with tactile cues (felt through the tool tip) to inform it about the larva’s depth in the hole and activity, and determine the most appropriate mode of manipulation of the tool. As usual, such work prompts more questions than it answers. We would like to know, for example, how juvenile crows acquire this behaviour and whether individuals differ in their ‘fishing’ tactics.

There are clearly still many open questions about tool use and manufacture in crows, and it offers a rich domain for collaborations with AI researchers. There are a number of areas in which such collaborations could offer valuable new insights, and we will elaborate on a selection of those areas below.

1. A thorough analysis of the requirements of the crows’ environment (section 2.1) would provide a framework in which we could evaluate crows’ tool-related behaviour. For example, can we categorise the variety of problems (related to tools) that the environment poses? Do the relevant parameters of these tasks vary dynamically in space and time? Which of these influence the crows’ evolutionary fitness, and which are irrelevant to its survival? If one was designing an agent to address the requirements of the environment (or the subset identified as important to evolutionary fitness), what kinds of representations would the agent require, and how would it process information, develop and learn? If the design space resulting from this process differs strikingly from the kinds of behaviours we know to exist in real crows, can we speculate about why evolution did not proceed down that path?
2. If the range of designs generated by the requirements analysis differs from range of behaviours shown by the crows, it might prompt us to investigate hitherto unnoticed competences, thus generating new hypotheses (see section 2.3). For example, given the difficulty of understanding the world from the perspective of a crow, it is possible that we have overlooked some subtle but important detail. Equally, if we test the generated hypotheses to the best of our ability and find that crows truly lack that competence (or form of representation, ability to apply knowledge in new contexts, etc.), then we have to attempt to explain its absence.
3. Adopting an integrated, systems approach (section 2.4) may help

us to make sense of how the varied components of the crow's cognitive system interact. For example, we currently make a number of possibly unwarranted assumptions about the information available to the animal, on which it might base decisions about how to solve a problem, because we do not have detailed knowledge of the sensory information available to the animal while it is manipulating a tool in a hole. Neither do we know what sort of ontology is available to the crow's perceptual mechanisms in interpreting that sensory information [22]. We need to consider the sensory, motor and information processing systems as a richly interconnected, overlapping and interacting set of processes which relate to, inform and modify one another.

## 4 MAKING PROGRESS

While there are ongoing collaborations between groups of biologists and AI researchers, conferences, workshops and symposia, such as the current one, are important in broadening participation, and providing a forum for discussion of the problems and their possible solutions. One important issue that we are aware of is that the two fields tend not to read the journals in which the other publishes. In addition, publishing in journals on the other side of the disciplinary boundary is also rare. As a consequence, researchers in each field are often unaware of advances made in the other, so opportunities to make advances are lost. Along with research collaborations, journals need to be more open to publishing papers from related fields, so that more people are exposed to new ideas which might spark collaborations. This might be a more fruitful route than attempting to get researchers to read new journals, since it is already a difficult task to keep up with the literature in one's own field.

One potential barrier to this process is the lack of a common set of terminology. In both fields, there are examples of different terms for the same concept or process, and examples where the same term is used in different ways in each field (for example, the term 'learning' has subtly different meanings in biology and AI, and there are probably more categories of 'learning' in biology). This is obviously a potential source of confusion and wasted effort, as researchers can end up talking at cross-purposes and misunderstanding one another's arguments. On a more optimistic note, such misunderstandings can provide a good opportunity to refine and clarify concepts, because explaining them to someone outside the field often reveals inconsistencies or previously overlooked gaps.

Finally to make progress with integrating AI and biology, we need to keep an open mind, and maintain realistic expectations about what can and cannot be achieved.

## 5 CONCLUSION

In this paper, we have attempted to explain the reasons why we think — from our perspective as biologists — that AI has the potential to make a significant contribution to the study of NI, by providing new conceptual frameworks and analytical tools. However, we would not propose such an interaction unless we believed that it would also involve benefits and scientifically interesting questions for AI researchers. After all, true altruism towards unrelated individuals is rare in nature — reciprocal altruism is more popular in social species! One such benefit involves the current role biology plays in AI, namely providing more detail of the competences of natural systems, and inspiration for artificial systems. In addition, studying the full range of NI in detail can clarify goals and assumptions in AI. For example, if you find species that can interact with the world in

a very flexible and sophisticated way, but lack language, does your autonomous robot necessarily need the ability to understand and produce language? If some animals lack trichromatic colour vision or binocularity, does a robot necessarily need either capability? The answer will depend on the intended 'niche' of the robot (and reveal another way that biologists can help), but thinking about these issues in relation to real animals may help to clarify existing assumptions.

Biology can also provide a new perspective on debates, such as those surrounding the importance and varieties of embodiment (see [37] for a recent review). The sensory and motor apparatus through which an animal interacts with the world has undeniably important effects on its cognition, as does the social or cultural environment in which it lives. However, convergent evolution has happened many times and ensures that — though two species may have radically different sensory, motor and information processing 'hardware' (think orangutans and octopuses) — they can emerge with striking similarities in their cognitive abilities ('software'). So while embodiment is important for the richness of the problem that it sets evolution, there are many possible solutions to the problem.

We hope that by interacting with AI researchers, we can also find a variety of new solutions to our various mutual problems.

## ACKNOWLEDGEMENTS

We would like to thank Aaron Sloman, Nick Hawes and Jeremy Wyatt for many fruitful and interesting discussions on this topic.

## REFERENCES

- [1] R.C. Barnett, R.P. Cole, and R.R. Miller, 'Temporal integration in second-order conditioning and sensory pre-conditioning', *Animal Learning and Behavior*, **25**, 221–233, (1997).
- [2] L.A. Bluff, J. Troscianko, A.A. Weir, A. Kacelnik, and C. Rutz, 'Tool use by wild New Caledonian crows *Corvus moneduloides* at natural foraging sites', *Proceedings of the Royal Society of London B*, (Published online 6 Jan 2010).
- [3] J. Bryson and L. Stein, 'Modularity and design in reactive intelligence', *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, (2001).
- [4] R.W. Byrne and L.A. Bates, 'Sociality, evolution and cognition', *Current Biology*, **17**, R714–R723, (2007).
- [5] J.G.H. Cant, 'Positional behaviour of female Bornean orangutans (*Pongo pygmaeus*)', *American Journal of Primatology*, **12**, 71–90, (1987).
- [6] J.G.H. Cant, 'Positional behavior and body size of arboreal primates: a theoretical frame-work for field studies and an illustration of its application', *American Journal of Physical Anthropology*, **88**, 273–283, (1992).
- [7] J. Chappell and A. Kacelnik, 'Selection of tool diameter by New Caledonian crows *Corvus moneduloides*', *Animal Cognition*, **7**, 121–127, (2004).
- [8] J. Chappell and A. Sloman, 'Natural and artificial meta-configured al-tricial information-processing systems', *International Journal of Un-conventional Computing*, **3**(3), 211–239, (2007).
- [9] J. Chappell, 'Avian cognition: Understanding tool use', *Current Biol-ogy*, **16**(7), R244–R245, (2006).
- [10] J. Chappell and A. Kacelnik, 'Tool selectivity in a non-mammal, the New Caledonian crow (*Corvus moneduloides*)', *Animal Cognition*, **5**, 71–78, (2002).
- [11] L. Chittka and J. Niven, 'Are bigger brains better?', *Current Biology*, **19**(21), R995–R1008, (2009).
- [12] D.C. Dennett, *Brainstorms. Philosophical essays on mind and psychol-ogy*, MIT Press Cambridge, MA, (1978).
- [13] G.R. Hunt, 'Tool use by the New Caledonian crow *Corvus monedu-loides* to obtain Cerambycidae from dead wood', *Emu*, **100**, 109–114, (2000).
- [14] G.R. Hunt and R.D. Gray, 'Direct observations of pandanus-tool man-ufacture and use by a New Caledonian crow (*Corvus moneduloides*)', *Animal Cognition*, **7**(2), 114–120, (2004).

- [15] G.R. Hunt and R.D. Gray, 'Species-wide manufacture of stick-type tools by New Caledonian crows', *Emu*, **102**(4), 349–353, (2002).
- [16] A. Karmiloff-Smith, 'Precis of beyond modularity - a developmental perspective on cognitive science', *Behavioral and Brain Sciences*, **17**(4), 693–706, (1994).
- [17] S.G. Larson, D. Schmitt, P. Lemelin, and M. Hamrick, 'Limb excursion during quadrupedal walking: how do primates compare to other mammals?', *Journal of Zoology*, **255**, 353–365, (2001).
- [18] J. McCarthy, 'The well-designed child', *Artificial Intelligence*, **172**(18), 2003–2014, (2008).
- [19] D.J. Povinelli and J. Vonk, 'Chimpanzee minds: suspiciously human?', *Trends in Cognitive Sciences*, **7**(4), 157–160, (2003).
- [20] D.M. Raup and A. Michelson, 'Theoretical morphology of the coiled shell', *Science*, **147**, 1294–1295, (1966).
- [21] D. Schmitt, 'Compliant walking in primates', *Journal of Zoology*, **247**, 149–160, (1999).
- [22] A. Sloman, 'Ontologies for baby animals and robots. From "baby stuff" to the world of adult science: Developmental AI from a Kantian viewpoint', Online tutorial presentation: <http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#brown>. (2009).
- [23] A. Sloman, 'The design-based approach to the study of mind (in humans, other animals, and machines), including the study of behaviour involving mental processes', Available at: <http://www.cs.bham.ac.uk/research/projects/cogaff/misc/design-based-approach.html>.
- [24] A. Sloman, 'What cognitive scientists need to know about virtual machines', in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, eds., N.A. Taatgen and H. van Rijn, pp. 1210–1215, (2009).
- [25] A. Sloman and J. Chappell, 'The altricial-precocial spectrum for robots', in *Proceedings IJCAI'05*, pp. 1187–1192, Edinburgh, (2005), Available at <http://www.cs.bham.ac.uk/research/cogaff/altricial-precocial.pdf>.
- [26] A. Sloman and J. Chappell, 'Computational cognitive epigenetics', *Behavioral and Brain Sciences*, **30**(4), 375–376, (2007).
- [27] A. Sloman, J. Wyatt, N. Hawes, J. Chappell, and G-J.M. Kruijff, 'Long term requirements for cognitive robotics', in *Proceedings CogRob2006, The Fifth International Cognitive Robotics Workshop. The AAAI-06 Workshop on Cognitive Robotics*, (2006).
- [28] E. Spelke, 'Initial knowledge: six suggestions', *Cognition*, **50**, 431–445, (1994).
- [29] S.K.S. Thorpe and R.H. Crompton, 'Locomotor ecology of wild orangutans (*Pongo pygmaeus abelii*) in the Gunung Leuser ecosystem, Sumatra, Indonesia: A multivariate analysis using log-linear modelling', *American Journal of Physical Anthropology*, **127**, 58–78, (2005).
- [30] S.K.S. Thorpe and R.H. Crompton, 'Orangutan positional behavior and the nature of arboreal locomotion in Hominoidea', *American Journal of Physical Anthropology*, **131**, 384–401, (2006).
- [31] S.K.S. Thorpe and R.H. Crompton, and R. McNeil Alexander, 'Orangutans utilise compliant branches to lower the energetic cost of locomotion', *Biology Letters*, **3**, 253–256, (2007).
- [32] S.K.S. Thorpe, R.L. Holder, and R.H. Crompton, 'Origin of human bipedalism as an adaptation for locomotion on flexible branches', *Science*, **316**, 1328–1331, (2007).
- [33] S.K.S. Thorpe, R. Holder, and R.H. Crompton, 'Orangutans employ unique strategies to control branch flexibility', *Proceedings of the National Academy of Sciences*, **106**(31), 12646–12651, (2009).
- [34] J. Troschianko, L.A. Bluff, and C. Rutz, 'Grass-stem tool use in New Caledonian Crows *Corvus moneduloides*', *Ardea*, **96**, 2, (2008).
- [35] A.A.S. Weir, J. Chappell, and A. Kacelnik, 'Shaping of hooks in New Caledonian crows', *Science*, **297**, 981, (2002).
- [36] A.A.S. Weir and A. Kacelnik, 'A New Caledonian crow (*Corvus moneduloides*) creatively re-designs tools by bending or unbending aluminium strips', *Animal Cognition*, **9**(4), 317–334, (2006).
- [37] T. Ziemke, 'Are robots embodied?', in *First International Workshop on Epigenetic Robotics*, volume 85, pp. 75–83. Lund University Cognitive Studies, (2001).

# On the practical nature of artificial qualia

Raúl Arrabales<sup>1</sup>, Agapito Ledezma<sup>1</sup> and Araceli Sanchis<sup>1</sup>

**Abstract.** Can machines ever have qualia? Can we build robots with inner worlds of subjective experience? Will qualia experienced by robots be comparable to subjective human experience? Is the young field of Machine Consciousness (MC) ready to answer these questions? In this paper, rather than trying to answer these questions directly, we argue that a formal definition, or at least a functional characterization, of artificial qualia is required in order to establish valid engineering principles for synthetic phenomenology (SP). Understanding what might be the differences, if any, between natural and artificial qualia is one of the first questions to be answered. Furthermore, if an interim and less ambitious definition of artificial qualia can be outlined, the corresponding model can be implemented and used to shed some light on the very nature of consciousness. In this work we explore current trends in MC and SP from the perspective of artificial qualia, attempting to identify key features that could contribute to a practical characterization of this concept. We focus specifically on potential implementations of artificial qualia as a means to provide a new interdisciplinary tool for research on natural and artificial cognition.

## 1 INTRODUCTION

It is well known that human conscious perception is not directly based on data acquired by senses, but heavily biased by psychological aspects like, for instance, cognitive context, subject history, and expectations. Additionally, context and subjective history is in turn shaped by the specific way in which stimuli are consciously perceived. Although we tend to believe that we perceive the reality, the fact is that qualia generated in our brains are far from being a truthful representation of real world. Nevertheless, generally our conscious experience of the world proves to be highly reliable in terms of everyday tasks. In short, the world is interpreted by the subject in such a way that is advantageous for his or her goals.

The former intuitive definition of conscious perception does not seem to be sufficient to build a comprehensive model of a conscious machine that includes phenomenal aspects (see for instance [1] for a discussion about the grand illusion of consciousness and perceptual phenomenology). We do not seem to have a satisfactory model or theory about what exactly qualia could be (either in conscious machines, humans or other animals). In other words, the phenomenal dimension of consciousness, both in natural and artificial creatures, remains elusive to scientific study. Additionally, as pointed out by Sloman [2], many times we might be discussing bogus concepts due to the use of misleading contexts and ill-defined terms. Although we are proposing here a set of new definitions, it is not

our aim to contribute to the existing confusion in the field of MC, but to help clarifying concepts from the engineering point of view. As argued by Sloman [2], directing the basic questions about consciousness to machines with different designs can help to figure out what really needs to be explained.

While many Artificial Intelligence (AI) implementations cover some aspects of the broader picture of cognition, we have not yet reached the point where human-like conscious machines are possible [3]. One of the greatest challenges that still needs to be addressed is the design of computational models of qualia, i.e. models of artificial qualia. MC designers are in the need of a practical definition of what could be an artificial conscious mind.

In this work, we assume that a complete and scientifically established definition of what artificial qualia could be does not yet exist. Therefore, we propose to circumvent this problem by looking for alternative, interim, and partial definitions that could contribute to the development of MC research field.

This approach does not necessarily lead to better implementations in terms of performance, but the whole exercise might at least provide more insight about what conscious machines could be like, what design strategies appear to be more promising, and how neuroscience could benefit from the application of computational models of qualia. As Chrisley has suggested [4], one of the aims of SP should be to characterize the phenomenal states possessed or modelled in MC implementations.

In the remainder of this paper we attempt to provide a decomposition of the artificial phenomenology problem into more tractable and recognizable steps. It is our hypothesis that the partial definitions of artificial qualia that we outline here may become useful conceptual tools in the domain of SP. Initially, in Section 2, we analyze the context of this research and introduce some of the constraints and difficulties that we have to deal with. Relevant related work is briefly discussed in Section 3. Section 4 covers the specific characterization of artificial qualia that we put forward. Then, in Section 5, the proposed approach is considered from the point of view of practical scientific applications. We illustrate this topic by suggesting that qualia could be studied in machines with relatively simple experiments involving, for instance, the apparent motion effects in visual experience. Finally, conclusions are presented in Section 6.

## 2 TACKLING THE PROBLEM

Taking into account the complexity of the problem described in the previous section, we have no other option but to bear with some of the well-known but controversial issues of the scientific study of consciousness. In relation with the specification of phenomenal states, it is convenient to distinguish between different components of such an intricate concept as consciousness. Additionally, the way in which phenomenology can be studied satisfactorily has to be analyzed. In particular, the problem of private first-person observations is discussed.

<sup>1</sup> Dept. of Computer Science, Carlos III University of Madrid, 28911. Madrid, Spain. Email: rarrabal@inf.uc3m.es.

## 2.1 Decomposing a complex concept

Many problems in the science of consciousness are rooted in the fact that the term consciousness can be used to refer to multiple related aspects [5]. In other words, consciousness can be seen as different phenomena according to the perspective of the observer. If we distinguish between phenomenal and functional dimensions of consciousness as suggested by Block [6], the brief description of the perception process outlined in Section 1 can be seen from two different perspectives:

- a) *Conscious perception is the set of phenomenal experiences of which our inner life is made of (the ‘what is it like’ to have experiential mental states [7]), e.g. the redness of red.*
- b) *Conscious perception is the set of functional representations or internal models of the world adapted to our needs, which are made available for use in reasoning and action, e.g. the neuronal encoding of color in the brain (see [8] for details).*

These two views (phenomenal consciousness and access consciousness respectively) should not be considered exclusive or contradictory, but complementary aspects of the same complex process. In fact, natural cognition shares these properties: the conscious contents of our minds are both experiential mental states and also functional representations accessible for reasoning and action. As Haikonen has suggested [9], qualia are the direct products of the perception process, and without qualia there is no consciousness. Therefore, qualia cannot be neglected in the study of consciousness, especially if we want our MC computational models to be of any use to the quest for the understanding of human cognition.

While qualia are usually associated with the first view (phenomenal consciousness), most of the work done in the domain of artificial cognitive systems is exclusively related to the second view (access consciousness). One of the reasons for this bias is the poor comprehension of phenomenal aspects of consciousness. Another significant reason is that plenty of work still needs to be done on machines that apparently do not require qualia for successful performance. A related problem is to determine why some machines might need qualia. These issues are briefly covered in the next subsections.

## 2.2 Avoiding the first-person/third-person problem

The problem of considering both personal and access views at the same time is that phenomenal consciousness is only available to the first-person observer, i.e. it is a private property [10]. Inner experience in human counterparts is usually inferred by third-person observers based on the similarity to their own case: if I feel pain when I get hurt, I infer other humans will likely feel the same in the same situations (because they have a nervous system like mine). However, when it comes to detecting the presence of phenomenal states in machines we do not even have the similarity argument as a factor to take into account in the inference process.

Being phenomenal qualia inherently private, how could we determine if a machine is experiencing any inner life? (See [11] for a comprehensive discussion about the problem of measuring

machine consciousness) This issue is quite related to the so-called hard problem of consciousness [12], which seems to remain an open problem. Essentially, we do not yet have any convincing explanation of phenomenal consciousness (at least, we do not have a theory that could be translated into a computational model).

Does this mean that any attempt to create artificial qualia will be futile? Or should we, instead of giving up the challenge, try to explore machine qualia as a means to shed light on the nature of consciousness? Do we really need to understand the very nature of consciousness in order to reproduce it in machines? Is there a real lack of scientific tools to address the problem of consciousness? Can we develop a model of phenomenal consciousness exclusively based on third-person approaches? Perhaps these questions cannot be answered yet; however, we believe third-person approaches can be successfully applied to make progress in the field of artificial cognitive systems and their application in AI-inspired biology.

Typically, applying third-person approaches involved looking exclusively at behaviour, including forms of accurate report [13]. However, external observers are also able to inspect the architecture and inner machinery of the creature. The inner inspection and monitoring of biological living organisms, including humans, is much more problematic than the inspection of working implementation of artificial cognitive systems. Therefore, the analysis of correlations between observed behaviour and internal inspection of MC implementations has to be exploited as it could provide valuable information about the models being tested (without the existing limitations in analogous experiments with biological creatures).

We argue that, following this line of research, a limited definition of artificial qualia can be made in such a way that only third-person approaches are used. This partial definition might not explain phenomenal qualia as it is present in humans, but it could be used to create computational models, and subsequent implementations, which then could be used to enhance our understanding of “natural consciousness”.

## 2.3 The function of qualia

Understanding what is the function of qualia and why they emerged as part of biological evolution is an essential part of the challenge of the scientific study of consciousness. As usual, the research interplay between natural and artificial sciences can be seen both ways: on one hand, a comprehensive understanding of qualia, as they manifest in biological creatures, might make possible the building of conscious machines; on the other hand, the path to a complete understanding of qualia in biology might lay through the research on new computational models focused on phenomenology.

These ideas about qualia are not free of controversies. While some argue that qualia are mere epiphenomena (e.g. [14]), we believe phenomenal consciousness appeared as an evolutionary advantage. One way to prove it would be to compare the performance of phenomenally conscious machines versus unconscious machines, both confronted to complex tasks in unstructured environments. Given that such an experiment is not realizable nowadays, we will focus on both evidence from the biological world and current computational models.

There are many features related with consciousness that we know are useful because they contribute to survival (for instance,

Theory of Mind [15], to name an illustrative one). These cognitive capabilities have a function and that is the reason why they have been selected by evolution. But, what about phenomenal aspects of consciousness? Do they have a clear functional role?

Qualia or subjective experience should not be seen as an additional component of the complex notion of consciousness, but a process that is present in relation with cognitive features. In the context of the proposed functional role, qualia are experienced by a creature when it is able to introspect some of its perceptual processes and use that introspection to generate a meta-representation which in turn is used to modulate the whole system. Qualia are indeed the output of the perception process [9], which in some cases are made explicit thanks to a transparent access to the perception process outcome (sensory system response to stimuli). In short, when we are aware of a red object in our field of view, we do not perceive the colour red, but the redness quale, which is the reaction of our perceptual system to the red colour stimulus.

The role of qualia described above could be studied in artificial systems. The generation of artificial qualia along these lines could provide insight about consciousness applicable to biological creatures. As argued by Sloman and Chrisley [16], a machine could even develop private ontologies for referring to its own private perceptual contents and states. The use of this ontology for modulating system processes is the function of qualia. A system with qualia is a system with *meta-management* capabilities (combination of introspection and active control based on self-monitoring). Making a serious effort to design and build such systems will contribute to the confirmation or refutation of these hypotheses about the role of qualia in biological creatures.

### 3 RELATED WORK

Neglecting the hard problem of consciousness has been one of the most common approaches in the practical design of artificial consciousness models. We think this position should not be criticized; at least when the phenomenal dimension is ignored explicitly, because tackling initially some of the so-called *easy problems* (see [12]) seems to be a plausible engineering strategy to begin with. However, as we pursue more ambitious objectives, like the ones outlined in the previous section, the field of MC should tend to adopt more tools based on the concept of Synthetic Phenomenology [4]. This aspect could be essential for the progress of existing research lines in the direction of human-like machines.

The analysis of the computational correlates of artificial qualia done by Chella and Gaglio is a remarkable example in which the authors have directly addressed the concept of artificial phenomenology [17]. In this work, based on the cognitive architecture developed by the same researchers [18], an active process integrating internal and external flows of information is designed to build a two-dimensional, viewer dependent, reconstruction of the subjective scene as perceived by the robot. It is argued that the matching process between this reconstruction and internal perceptual data represents the artificial qualia of the system.

Haikonen's cognitive architecture is another example of a computational model that takes into account the generation of artificial qualia [9, 19]. It is argued that a realization of the

conscious machine proposed by Haikonen would be endowed with an inherent mechanism to produce qualia with grounded meanings. Additionally, the machine would be able to report these qualia via secondary symbols such as uttered words.

Even though other MC implementations have not been initially designed to have phenomenal consciousness, their capacity for sustaining phenomenal states can be explored. This is for instance the case of LIDA [20], which is based on the Global Workspace Theory [21], and has been designed to be functionally (but not phenomenally) conscious. What mechanisms would need to be added to an implementation like LIDA in order to render her phenomenally conscious (i.e., in order to have her generating and reporting artificial qualia)? As pointed out by her creators [22], implementing a mechanism for the generation of a stable, coherent perceptual world, along the lines discussed by Merker [23, 24] (i.e., suppressing apparent motion produced by the movement of sensory receptors) might contribute to the design of phenomenally conscious machines.

Another significant work towards the definition of human-like information processing architectures is H-CogAff [25], where a meta-management layer based on reflective processes is added. This particular case of CogAff [26] is aimed at specifying a minimal architecture for a human-like system. In this work the issue of qualia is addressed as an architecture-based concept [16]. More specifically, virtual machine architectures are considered as a suitable domain for current MC analyses and experimentation. The virtual machine functionalism provides an account for phenomenology in which virtual machines, although not being physical, are real machines that can affect and be affected by the physical environment (see [16] for a thorough discussion on this point). Moving the discussion about qualia to the level of description of virtual machines – which is not definable in terms of physical sciences – might help clarifying what qualia actually are.

In relation with the problem of testing for the presence of artificial qualia, Gamez has proposed a scale that attempts to estimate the probability of a machine being able to sustain phenomenal states [27]. This scale makes no assumptions as to what specific mechanisms could be used to generate qualia. It is, instead, based on the similarity to human brain in terms of complexity.

More elaborated measures have been proposed to detect the presence of phenomenal states in neural systems. According to Tononi's theory of information integration [28], phenomenal consciousness can be detected looking at the level of dynamical complexity of a system, i.e. the combination of integration (unity of conscious experience) and differentiation (ability to discriminate conscious experiences amongst a vast repertoire of possible scenes). According to this, an artificial system would be able to experience qualia only if it is able to sustain a high level of information integration.

Apart to its theoretical application to biological organisms, Aleksander and Gamez have used Tononi's information integration theory to analyze how different artificial neural network topologies generate different effective information levels [29]. This work is focused in the discriminative or informational value of qualia. Although information integration can be considered a fundamental aspect in relation with qualia, other properties, like the ones discussed above, have to be taken into consideration [11].

## 4 CHARACTERIZING ARTIFICIAL QUALIA

In the domain of MC research, designers have to deal with the concept of qualia in order to develop implementations that could be claimed to be conscious. One way to mitigate the complexity involved in this task is to conceptually decompose the notion of qualia into different aspects that can be analyzed separately. It is our hypothesis that working with these partial views might provide some insight to the whole picture. What follows is a specific proposal of partial but complementary definitions of artificial qualia.

### 4.1 Partial definitions of artificial qualia

We propose to distinguish between three different stages or characterizations of the development of mechanisms that support qualia in machines (see Fig. 1 for an illustration of the relation between these stages; see Section 5 for an explanation of the content represented in Fig. 1):

- **Stage 1. Perceptual Content Representation.** At this stage the information acquired by the perceptual system of the machine is integrated and interpreted, generating a subjective representation. This content is built as a result of the combination of exteroceptive and proprioceptive sensing subsystems, hence giving place to an inherently subjective content representation. The process that generates this perceptual content involves a continuous checking for consistency. In other words, a number of possible partial reconstructions of the world compete for being integrated into the final consistent match. This match, or inner world final reconstruction, is achieved by a coherent integration between what is being perceived from the external world and what it currently represented as inner depiction (see Dennett's Multiple Draft Model [30], for a more metaphorical description of these sorts of competitive/collaborative content creation process).
- **Stage 2. Introspective Perceptual Meta-Representation.** This stage refers to the monitoring of the processes mentioned in stage 1, and also the creation of derived meta-representations. Observing how machine's own perceptual content is created can potentially result in a private ontology (meta-representation) about what is it like for the machine to experience subjective perceptual contents.
- **Stage 3. Self-Modulation and Reportability.** In the case of machine being able to achieve stages 1 and 2, meta-representations from stage 2, or introspective ontologies [2], could be used to modulate the way all perceptual systems work (including stages 1 and 2). This constitutes a self-regulation loop that has clear functional implications; i.e. qualia as defined here is part of a causal process. Additionally, introspective ontologies created in stage 2 could be used to report the artificial qualia of the machine.

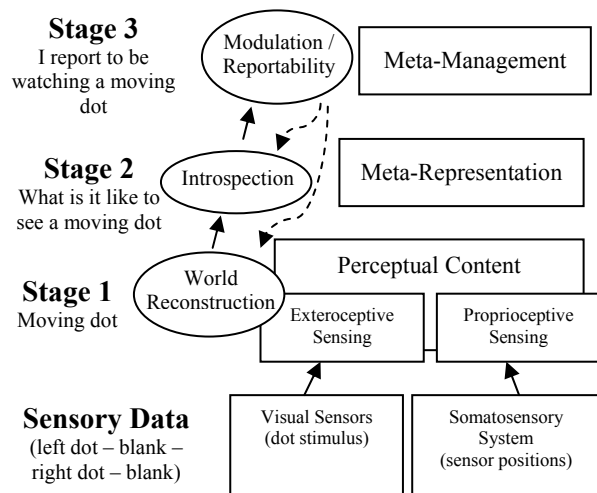


Figure 1. Stages in the development of artificial qualia

Stages described above make no claims about the qualities or modalities of the specific contents of the artificial mind, or to what extent they could resemble human subjective experience. The presence of different sensory modalities (laser ranging, for instance) and different mechanisms for cognition will produce different conscious contents and associated qualities. Nevertheless, we argue that the abstract mechanisms described above resemble those that generate qualia in biological organisms. We suggest that this claim could be supported by the application of the proposed framework in computational models, thus enabling a comparative analysis between qualia generated by humans and qualia generated in machines.

The proposed stages are just the components of a conceptual framework or guideline for the design of MC architectures. The subsequent implementations and experiments are expected to clarify some aspects of the nature of phenomenal experience and its impact in cognitive abilities.

Self-consciousness is not specifically addressed in the proposed definition because it is not considered a requirement for phenomenal consciousness. Nevertheless, self-consciousness could be explained in the context of the proposed framework having a model of the body as part of the perceptual content representation. The concept of self would be expected to arise as a meta-representation in stage 2. Then, references to the self could be found in accurate reports generated in stage 3.

### 4.2 Detecting the presence of qualia

The definition of qualia that we have described represents a hypothesis to be tested. To be precise, it should not be taken as a valid and established knowledge. That would lead us to think that the presence of qualia can be scientifically tested just by detecting the proposed mechanisms through inspection. In contrast, what we suggest is to use this approximation to what qualia might be as a working hypothesis. Thus approach calls for the experimentation with MC implementations designed following these assumptions.

The results of this testing process are expected to prove whether or not the original hypothesis was correct. One of the

benefits of this kind of approach is that the first-person problem mentioned above can be circumvented. Nevertheless, more work should be done in order to design meaningful experiments combining both behavioural outcome and architecture inspection (see for instance [31] for a particular discussion of misleading experiment design about self-consciousness using the mirror test). Also, identification of hallmarks of qualia, like bistable perception [32], should be put in the context of the present discussion.

## 5 VISUAL EXPERIENCE AND APPARENT MOTION

In order to illustrate the proposed segmented characterization of artificial qualia, the conscious perception of apparent motion is analyzed. Humans can perceive motion not only from real moving objects, but also from series of images containing spatially segregated visual stimuli [33]. Simple experiments to test this consist of two stationary blinking stimuli which are presented using different spatial and temporal parameters (see Fig. 2). At certain rates, subjects perceive motion (apparent motion).



**Figure 2.** Sequence of images used to generate apparent motion qualia in humans

The sequence of images depicted in Fig. 2 are used to generate apparent motion qualia in humans, and presumably also in phenomenally conscious machines. Note that a blank inter-stimulus interval (ISI) is inserted after every dot stimulus (the looping sequence is as follows: left dot stimulus – blank – right dot stimulus – blank).

This experiment, which is usually carried out with human subjects, could be carried out using a machine as subject. Putting this experiment in the context of the former characterization of qualia, let us consider a robot with a visual perceptual system modelled after the human visual cortex. The basic content at each stage could be described as:

- Stage 1: *“moving dot”*.
- Stage 2: *“what is it like to see a moving dot”*.
- Stage 3: *“I report to be watching a moving dot”*.

Fig. 1 represents these different levels of content in each of the stages that we have defined for the development of artificial qualia.

The perception process in the robot would follow these steps: first of all, visual sensors, let's say a stereo visual system composed of two digital cameras, acquires the images using their light detection sensors. At the same time, robot's somatosensory system acquires the relative position of the cameras, their orientation, and their foveation. The combination of sensory data from exteroceptive sensors (pixel maps from cameras) and proprioceptive sensors (relative coordinates from camera

position and focus sensors) is then used to form depictions of percepts along the lines described by Aleksander and Dunmall [34].

As the sequence described in Fig. 2 is presented to the robot, single depictive percepts are created to represent the appearance of dots and their relative positions. Subsequently, robot motion detectors (these detectors could be built for instance with artificial neural networks), fed with the stream of dot percepts, will eventually create new motion percepts depending, amongst other things, on the duration of ISI frames. These moving dot percepts (or moving dot representations) are the contents of *Stage 1* (*“moving dot”*).

The presence of the motion percepts will in turn trigger a set of reactions in the system. For instance, if the robot is designed to keep track of some moving objects, or detect some types of trajectories, the associated detectors will be activated. Also, affective evaluations of percepts could be invoked (the robot could be designed, or could have learnt, that moving dots have to be evaluated positively, and therefore maintain bonds with them). If the robot were endowed with a mechanism to represent these reactions, he would generate meta-representations of *“what is like”* for the robot to see a moving dot. This content corresponds to our *Stage 2* definition.

Finally, if *Stage 2* introspective content is used both for self-regulating global perception-action processes, and also for reporting purposes, then the robot would be able to reason explicitly about what does it mean to him to see a moving dot. Provided with the necessary linguistic skills, the robot will also be able to report his mental content using his own ontology (*Stage 3* content).

## 6 CONCLUSIONS AND FUTURE WORK

A developmental view of qualia based on former work by other authors has been defined as an attempt to provide a conceptual framework for the creation of new MC models. The definition of qualia presented in this paper advocates for a functional role of phenomenal consciousness. Furthermore, self-modulation and integration of perceptual systems is appointed as a process being driven by the construction of introspective meta-representations. Additionally, reportability is assumed to be based on the same meta-representations.

It is expected that implementations based on the proposed conceptual framework are able to incorporate the phenomenal dimension of consciousness into their models. That is not to say that they will become phenomenally conscious just because the proposed conceptual definition is considered. Nevertheless, we argue that exploring the design space in the proposed direction might shed some light to the problem of production of qualia, both in natural and artificial creatures.

We are currently working in the development of visual illusion experiments, like the one described in the previous section, using the cognitive architecture CERA/CRANIUM [35]. We expect to draw some significant conclusions in the domain of synthetic phenomenology from these sorts of experiments.

Phenomenology is typically one of the fields where advancement in natural sciences is more difficult and we think this discipline could benefit more from the research with artificial systems. It is especially in more challenging aspect of natural sciences where we think that using AI-inspired biology can sometimes provide some useful clues to boost research.

## ACKNOWLEDGEMENTS

This work was supported in part by the Spanish Ministry of Education under CICYT grant TRA2007-67374-C02-02.

## REFERENCES

- [1] A. Noë, "Is the Visual World a Grand Illusion?" *Journal of Consciousness Studies*, vol. 9, pp. 1-12, 2002.
- [2] A. Sloman, "Why some machines may need qualia and how they can have them: Including a demanding new turing test for robot philosophers." in *AI and Consciousness: Theoretical Foundations and Current Approaches - AAAI Fall Simposia 2007*, pp. 9-16.
- [3] R. Arrabales, A. Ledezma and A. Sanchis. Establishing a roadmap and metrics for conscious machines development. In *Proceedings of the 8th IEEE International Conference on Cognitive Informatics*. G. Baciú, Y. Wang, Y.Y. Yao, W. Kinsner, K. Chan & L.A. Zadeh (Eds.). Hong Kong. pp. 94-101. June 2009.
- [4] R. Chrisley, "Synthetic Phenomenology," *International Journal of Machine Consciousness*, vol. 1, pp. 53-70, 2009.
- [5] N. Block, "Some Concepts of Consciousness," 9, Oct. 20, 2001. Available: <http://www.nyu.edu/gsas/dept/philo/faculty/block/>
- [6] N. Block, "On a Confusion about a Function of Consciousness," *Behav. Brain Sci.*, pp. 227-287, 1995.
- [7] T. Nagel, "What Is It Like To Be a Bat?" *The Philosophical Review*, vol. 83, pp. 435-450, 1974.
- [8] P. Lennie, "The physiology of color vision," in *The Science of Color (Second Edition)* Steven K. Shevell, Ed. 2003.
- [9] P. O. A. Haikonen, "Qualia and conscious machines". *International Journal of Machine Consciousness* 1(2), pp. 225-234. 2009.
- [10] D. C. Dennett, "Quining qualia," *Consciousness in Modern Science*. 1988. Available: <http://cogprints.org/254/>
- [11] R. Arrabales, A. Ledezma and A. Sanchis, "Strategies for Measuring Machine Consciousness," *International Journal of Machine Consciousness*, vol. 1, pp. 193-201, 2009.
- [12] D. Chalmers, "Facing Up to the Problem of Consciousness," *Journal of Consciousness Studies*, vol. 2, pp. 200-219, 1995.
- [13] A. Seth, B. Baars and D. Edelman, "Criteria for consciousness in humans and other mammals," *Consciousness and Cognition*, vol. 14, pp. 119-139, 3. 2005.
- [14] F. Jackson, "Epiphenomenal Qualia," *The Philosophical Quarterly*, pp. 127-136, 1982.
- [15] L. S. Vygotsky, "Mind in Society: The Development of Higher Psychological Processes." 1980.
- [16] A. Sloman and R. Chrisley "Virtual machines and consciousness". *Journal of Consciousness Studies*, 10. pp. 133-172. 2003.
- [17] A. Chella and S. Gaglio, "In search of computational correlates of artificial qualia," in *Advances in Intelligent Systems Research (AGI-09)*, May 2009.
- [18] A. Chella, M. Frixione and S. Gaglio, "A cognitive architecture for robot self-consciousness," *Artif. Intell. Med.* 44(2), pp. 147-154. 2008.
- [19] P. O. A. Haikonen, *Robot Brains. Circuits and Systems for Conscious Machines*. UK: John Wiley & Sons, 2007.
- [20] S. Franklin, U. Ramamurthy, S. K. DiMello, L. McCauley, A. Negatu, R. L. Silva and V. Datla, "LIDA: A computational model of global workspace theory and developmental learning," in *AAAI Fall Symposium on AI and Consciousness: Theoretical Foundations and Current Approaches*, 2007, pp. 61-66.
- [21] B. J. Baars, *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press, 1988,
- [22] U. Ramamurthy, "Might a LIDA controlled robot be phenomenally conscious?" in *Proceedings of the Nokia Workshop on Machine Consciousness*, 2008, pp. 32-33.
- [23] B. Merker, "The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution." *Conscious. Cogn.* 14(1), 2005. pp. 89-114.
- [24] S. Franklin, "Evolutionary pressures and a stable world for animals and robots: A commentary on Merker," *Conscious. Cogn.*, vol. 14, pp. 115-118, 2005.
- [25] A. Sloman and M. Scheutz, "A framework for comparing agent architectures." Presented at In UKCI'02: Proceedings of the UK Workshop on Computational Intelligence. 2002.
- [26] A. Sloman, "Varieties of affect and the CogAff architecture schema." Presented at Proceedings Symposium on Emotion, Cognition, and Affective Computing AISB'01 Convention. 2001.
- [27] D. Gamez, "An ordinal probability scale for synthetic phenomenology," in *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness* .2005, pp. 85-94.
- [28] G. Tononi, "An information integration theory of consciousness." *BMC Neuroscience* 5(1), 2004. pp. 42.
- [29] I. Aleksander and D. Gamez, "Iconic training and effective information: Evaluating meaning in discrete neural networks," in *Biologically Inspired Cognitive Architectures II. Papers from the AAAI Fall Symposium*. AAAI Technical Report FS-09-01 2009, pp. 2-10.
- [30] D. C. Dennett, *Consciousness Explained*. Boston: Little, Brown and Co, 1991,
- [31] P. O. A. Haikonen, "Reflections of consciousness: The mirror test," in *Proceedings of the 2007 AAAI Fall Symposium on Consciousness and Artificial Intelligence*, 2007, pp. 67-71.
- [32] N. Fürstenau, "A computational model of bistable perception-attention dynamics with long range correlations". *KI 2007: Advances in Artificial Intelligence* pp. 251-263.
- [33] L. Muckli, N. Kriegeskorte, H. Lanfermann, F. E. Zanella, W. Singer and R. Goebel. (2002, May 1). Apparent motion: Event-related functional magnetic resonance imaging of perceptual switches and states. *J. Neurosci.* 22(9), pp. 219.
- [34] I. Aleksander and B. Dunmall, "Axioms and Tests for the Presence of Minimal Consciousness in Agents," *Journal of Consciousness Studies*, vol. 10, 2003.
- [35] R. Arrabales, A. Ledezma and A. Sanchis, "CERA-CRANIUM: A test bed for machine consciousness research," in *International Workshop on Machine Consciousness 2009*. Hong Kong. June 2009. pp. 105.

# Route Learning Through Classification

Bart Baddeley and Paul Graham and Andrew Philippides and Philip Husbands<sup>1</sup>

**Abstract.** A simple approach to route following is to scan the environment and move in the direction that appears most familiar. In this paper we investigate whether an approach such as this could provide a model of visually guided route learning in ants. As a proxy for familiarity we use the learning algorithm Adaboost [6] with simple Haar-like features to classify views as either part of a learned route or not. We show the feasibility of our approach as a model of ant-like route acquisition by learning a non-trivial route through a real-world environment using a large gantry robot equipped with a panoramic camera.

## 1 Introduction

Individual ant foragers show remarkable navigational performance, rapidly learning long idiosyncratic routes through cluttered environments [2], guided by learnt visual landmark information [16, 10, 4, 14, 5]. Studies of such visual navigation have revealed how insects combine simple strategies to produce robust behaviour and insect navigation is now an established model system for investigations of the sensory, cognitive and behavioural strategies that enable small-brained animals to learn and utilise complex sequences of behaviour in the real world.

One elegant use of visual landmark information that is part of the insect's navigational toolkit is view-based homing. Behavioural experiments with ants [15, 5] and bees [1] have shown that individuals store 2D retinotopic views of the world as seen from their goal location. Subsequent search for that goal location can be driven by a comparison of their current view of the world and the view stored from the goal location. As this is an efficient and economical way of pin-pointing a location, it isn't a great leap to imagine that knowledge of the world over larger scales, such as routes, could be internalised as a series of stored views that are linked together as a sequence. Route behaviour in this framework would entail homing from one stored view to another. However, recent studies with ants suggest that guidance along routes might not be best served by chains of snapshots. Behavioural experiments suggest that routes can be performed using simpler procedural rules where the rule governing a path segment can be associated with the appropriate visually identified location [3]. Moreover, attempts to model route behaviours using linked view-based homing have shown it to be a non-trivial problem which requires the agent to robustly determine at which point a waypoint should be set during route construction, and deciding when a waypoint has been reached during navigation [11]. Essentially, for robust route navigation, an agent therefore needs place recognition to determine where along the route it is [12]. In conjunction with environmental noise, these problems make robust route navigation a non-trivial aim.

We begin our study of visually guided routes by drawing a line under previous modelling which defines routes in terms of discrete waypoints. Instead, we define a minimal route learning process in which the route is learnt more holistically. Rather than learning and homing to a set of positions, the agent instead learns a more general mapping which associates the current view *not* with a particular place but instead with a particular action. For an ant or robot that can only translate in one direction relative to its body axis and has a fixed viewing direction, the direction of movement is determined by the viewing direction and visa versa. Thus, if the current retinotopic view is similar to a remembered view, it is likely that the current viewing direction is the also the correct direction to move in. Having constraints on movement direction and viewing direction means that a single view can define not only a location but a direction that should be taken at that place. Crucially however, we do not attempt to learn every view along the route, but instead use them to learn a classifier that can be applied to any view to determine how likely it is that it is part of the route. We suggest this approach is a powerful platform for investigating efficient encoding of route guidance information and how this depends on the visual ecology within which an agent navigates.

This tight coupling of sensation and action allows us to reframe the problem of navigation in terms of a search for the views that are associated with a route. By visually scanning the environment and moving in the direction that is most similar to the views encountered during learning an ant or robot should be able to reliably retrace a given route. Both desert ants and wood ants perform scanning behaviours that would support this approach. When released in an unexpected but familiar place the desert ant *melophorus bagoti* scan the environment by turning rapidly on the spot. More than one scan maybe performed with short straight runs of a few centimetres separating them before the ant finally sets off in a seemingly purposeful manner. Wood ants exhibit a second form of scanning behaviour. Instead of walking in a straight line, wood ants instead tend to weave a somewhat sinuous path. This has the effect of producing scans of the world centred on the overall direction of movement.

We propose that if ants are able to somehow recognise familiar views, then they can recapitulate routes by simply scanning the environment and moving in the direction that is deemed most similar to the views experienced during learning. Ants could of course simply remember the complete set of views that were experienced during learning, however this approach would result in an extremely high memory load. Instead we propose an approach that involves implicitly modelling the distribution of the views experienced during learning by using a classifier to determine whether a given view comes from part of the learned route or not. Using a classifier provides us with a more compact way of storing the information required to recognise familiar views. Crucially, the approach we employ also provides a measure of the expected uncertainty of the classification

---

<sup>1</sup> University of Sussex, UK, email: bartbaddeley@googlemail.com

(as will be explained later). Here we test the idea of using the familiarity of views as a means of navigation by training a classifier to determine whether a given view is part of the route or not and then using the learned classifier to dictate the direction of movement during route following.

Our results implemented on a real robot indicate that this is indeed a feasible navigational strategy allowing the learning of complex routes through cluttered environments such as might be experienced by a navigating ant.

## 2 Methods



**Figure 1.** The gantry robot used in all of the experiments.

To test our hypothesis we need to sample the world from an ant's view point. To do this we used a large volume Cartesian XYZ robot to sample panoramic images along a pre-specified ground-level trajectory through a cluttered environment [see Figures 1, 2 and 4].

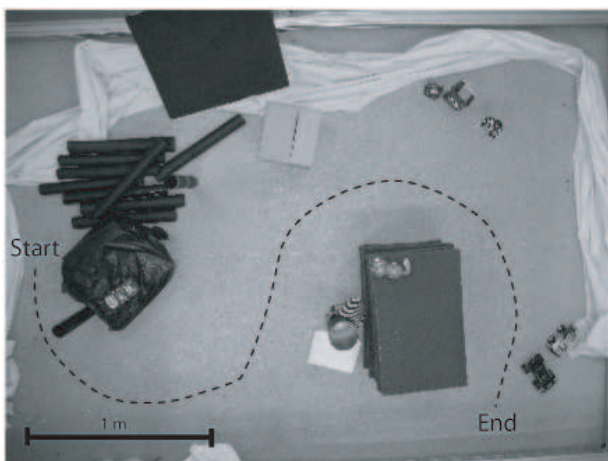


**Figure 2.** The environment viewed from above showing the learnt route (dotted line) through a cluttered environment. The scale bar indicates 1m.

In order to train a classifier it is necessary to generate positive and negative training examples of the input to be classified. In our

case this means collecting views that are part of the route and views that are not part of the route. The positive examples are simply the forward facing views experienced along the route. The negative views consisted of views from the route taken facing to the left and right of the direction of movement at an angle of  $\pm 45^\circ$  relative to the route heading. A small amount of normally distributed noise ( $s.d. = 0.1 \text{ radians}$ ) was added to each of the sampling directions. This approach is inspired by the observation that ants tend not to move in a straight line on a route but instead proceed in a sinous manner that results in some views that do not relate to the overall direction of travel and some that do [see Figure 3].

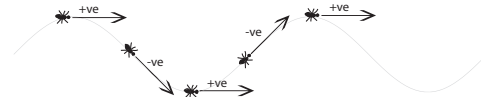


**Figure 3.** By weaving back and forth during locomotion it is possible to sample both the on-route and off-route views need to train a classifier.

Panoramic images were unwrapped and downsampled before performing feature extraction using simple Haar-like features [Figure 5]. Examples of the unwrapped panoramic images are given in figure 4 which shows an unwrapped image (top) and a heavily downsampled version of the same image (middle) together with the representation of the downsampled image in terms of the activations of a set of 100 Haar-like feature detectors (bottom).

Classifying images is a difficult task due to the high dimensionality of the input if we adopt a pixel by pixel representation. In order to make learning tractable we need to project this high dimensional space into a lower dimensional space that retains enough of the necessary structure to allow successful classification of the input. As a first step in reducing the dimensionality of the input we downsample the images to a resolution of ten degrees per pixel [Figure 4 middle]. To further reduce the dimensionality of the input we use simple Haar-like features [Figure 5] to construct a lower dimensional representation of each image. Each feature produces a single real valued output for a given image so we can control the dimensionality of the input to our navigational algorithm by defining the number of features that we use. In the current work we chose to use one hundred features selected from a randomly initialised pool of ten thousand. Finally we use the thresholded outputs of the features as simple classifiers. This leaves us with the problem of selecting the features from our pool of ten thousand and determining how to combine their outputs to form a reliable classification. Thankfully, boosting provides us with an approach that achieves both of these requirements.

### 2.1 Boosting

Boosting is a supervised learning technique for constructing a *strong classifier* from a set of *weak classifiers* given a training set of labelled positive and negative examples. A *weak classifier* is a classifier that performs only slightly better than chance. Conversely, a *strong classifier* is one that performs arbitrarily well. A *strong classifier* is constructed from a linear weighted combination of the outputs of weak classifiers.

There exist many variants of boosting algorithms. Adaboost [6], the approach we use in this paper is one of the most commonly used. The basic algorithm works as follows. At each iteration, the training data are resampled or reweighted according to a distribution of weights that indicate the current importance of each example

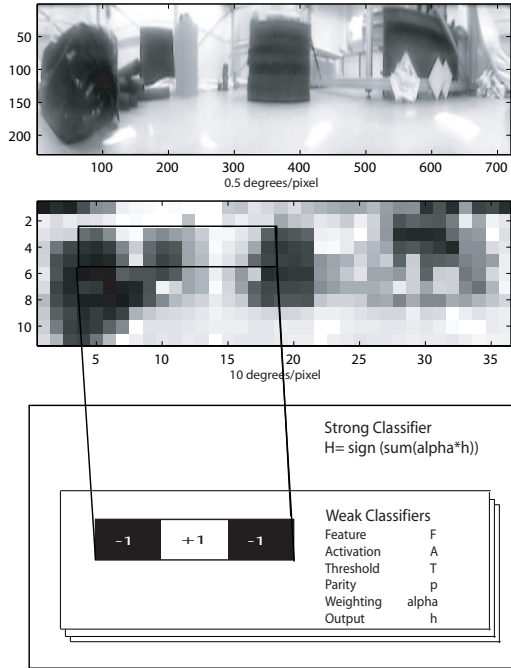


**Figure 4.** High (top) and low (middle) resolution panoramic images of a typical view from the workspace. The downsampled image is encoded using a set of one hundred Haar-like basis functions (bottom) that are fixed in space relative to the direction of view.

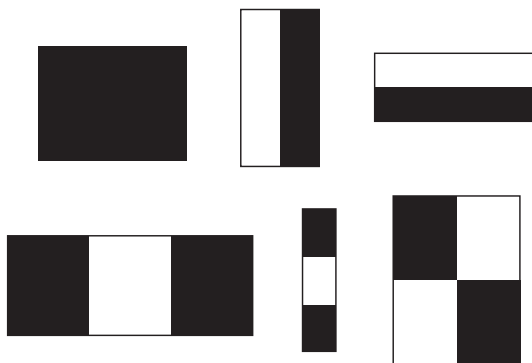


**Figure 5.** Examples of the six classes of Haar-like features that were used to represent the images. Each feature is defined in terms of a position a size and a class.

in the dataset. A *weak classifier* is then learned using this resampled/reweighted dataset and is added to the *strong classifier*. The relative contribution of each of the *weak classifiers* to the final *strong classifier* is determined by performance on the sampled data. Finally, the weights of incorrectly classified examples are increased and correctly classified examples decreased, thereby encouraging the next *weak classifier* to focus more on the examples that were incorrectly classified at the last iteration. *Weak classifiers* are added until the overall classification performance exceeds some threshold or the maximum number of *weak learners* is reached.

The pseudocode for Adaboost is as follows:

Set  $T$  = maximum number of weak classifiers

Given:  $(x_1, y_1), \dots, (x_m, y_m)$

where  $x_i \in X, y_i \in Y = \{-1, +1\}$

Initialize  $W_1(i) = \frac{1}{m}, i = 1, \dots, m$

For  $t = 1, \dots, T$  :

Find the classifier  $h_t : X \rightarrow \{-1, +1\}$  that minimizes the error with respect to the distribution  $W_t$

Choose  $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$  where  $\epsilon_t$  is the weighted error rate of classifier  $h_t$  with respect to the reweighted data.

Update  $W_{t+1}(i) = \frac{W_t(i) \exp(-\alpha_t \cdot y_i \cdot h_t(x_i))}{Z_t}$  where  $Z_t$  is a normalization factor that ensures that  $W$  represents a probability distribution over the training data

Output the final classifier  $H(x) = \text{sign} \left( \sum_{t=1}^T \alpha_t h_t(x) \right)$

Following Viola and Jones [13] we implement adaboost using single Haar-like features [Figure 5] as the basis of our *weak classifiers*. The Haar-like features consist of randomly chosen rectangular patches of the image which are then subdivided in one of four ways. The value of a *one rectangle feature* is simply the mean intensity value of the patch. The value of a *two rectangle feature* is the difference between the mean intensity of two rectangular regions. The regions have the same size and shape and are located next to each other either horizontally or vertically. The value of a *three rectangle feature* is given by the mean intensities of two outer rectangles subtracted from the mean intensity of a central rectangle, again oriented either horizontally or vertically. Lastly, a *four rectangle feature* computes the difference between diagonal pairs of rectangles. A weak classifier  $h_j(x)$  thus consists of a Haar feature  $f_j$ , a threshold  $\theta_j$  and a parity  $p_j$  that determines whether the output of the feature detector  $f_j$  should be greater than or less than the threshold  $\theta_j$ .

$$h_j(x) = p_j f_j(x) < p_j \theta_j$$

By providing a pool of feature detectors, each defining a weak learner, adaboost is able to perform feature selection. At each iteration a single feature detector is chosen that best aids in classification. This allows adaboost to pick out and use only those features that are most useful for the current classification problem.

Key to our use of a boosted classifier is the fact that it is possible to obtain a confidence value associated with any given classification made using the trained classifier. This confidence value is related to the margin and is given by:

$$\text{conf} = \left\| \sum_{t=1}^T \alpha_t h_t(x) \right\|$$

Which is simply the degree to which the the combined weak classifiers differ from zero, prior to the sign being taken.

By applying the classifier to views in different directions we can attempt to determine which of the views are from the learned route. By weighting each of the viewing directions that produce positive classifications by their associated confidence values we can determine a direction to move that is most likely to keep us on the learned route.

## 2.2 Data collection

All experiments reported here were performed on a gantry robot - a large volume XYZ Cartesian robot. The gantry axis configuration provides an operating volume of 3000 mm X 2000 mm X 2000 mm. The sensor end of the Z-axis can be placed anywhere within this volume with sub-millimetre accuracy. For the experiments presented here a catadioptric camera system (VCAM 360) is mounted on the Z-axis to produce panoramic images. A panoramic mirror projects a 360° image of the environment onto a downward facing CCD video camera. The image is transformed from a circular reflection into a panoramic image that is used for subsequent processing.

The gantry workspace was populated with a variety of objects consisting of foam blocks, piles of fabric, paper rolls and a random selection of toys. Objects were placed in such a way that it was possible to move the sensor head along a route through this visual clutter. Routes could be made more or less challenging by varying the degree of clutter and the straightness of the routes. In order to go beyond what is possible with a snapshot type model, the beginning and end points of all routes were chosen so that it would not be possible to perform the route using this approach. This is achieved by making sure that the end point of the route could not be viewed from the starting position.

## 2.3 Route learning through classification

Haar-like features were extracted from the set of training images and used to train a boosted classifier. During testing the camera was positioned at the start of the route facing in the correct direction. From this position images were sampled in a range of directions from  $-60^\circ$  to  $+60^\circ$  in steps of  $5^\circ$  relative to the current heading. Features were extracted from all of these images and used as input to the classifier. All of the viewing directions that produced a positive classification contributed to a weighted average with the weighting controlled by the confidence interval of the individual classifications. The weighted average was then used to determine the direction of travel and a 5cm step was made in this direction. The process was then iterated until success or failure.

## 3 Results

To test our approach a classifier was trained using a set of images gathered during a single traversal of the route shown in figure 2. The training data consisted of 171 views from 57 positions along the pre-specified route, 57 forward facing views and 57 views to both the left and right resulting in 57 positive and 114 negative views. A boosted classifier with 100 weak classifiers was trained on this dataset. Performance was then assessed by starting the robot in a series of different positions close to the original starting position. Figure 6 (top) shows the performance of the approach when starting at the same point as during the training run. Figure 6 (bottom) shows ten separate runs with starting positions at varying distances from the original start point.

The three leftmost starts fail due to the path exiting the permissible workspace of the robot. The fourth leftmost start fails having successfully negotiated the first corner suggesting that the previous three starts would also have failed had they not stopped due to leaving the workspace. The remaining six starts including the original start position (position five) all successfully recapitulate the original route.

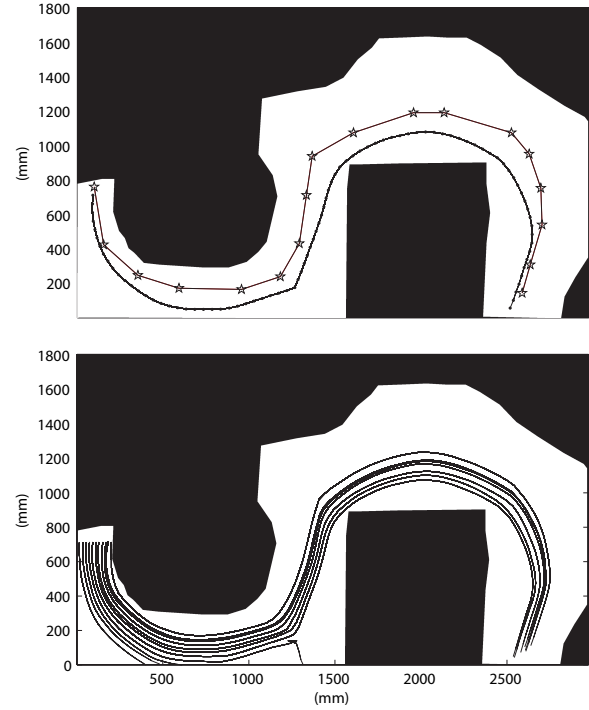


**Figure 6.** Performance of the algorithm. Top) Training data were collected along the route indicated by the linked stars and the resultant learned route is indicated by the dots. Note the symbols on the two routes do not indicate the points at which images were taken although for the training route they do indicate the waypoints that defined the route. Bottom) Ten example runs of the navigation algorithm. The ten starting positions are centred on the original starting position. Six out of ten trials are successful. Black regions indicate areas of the workspace that were defined as out of bounds due to potential damage to the camera from collisions with objects.

In order to get an understanding of how the algorithm would perform across the entire environment. We moved the panoramic camera across a grid covering the entire workspace. At each location we scanned in all directions and used the classifier to determine a preferred direction of movement, together with a confidence in its prediction. The results of this analysis are shown in figure 7 with confidence indicated by the length of the arrows. Contrary to what we would expect from the performance observed during route following, when input is sampled from all directions at all positions many views are clearly erroneously classified as being associated with the route as evidenced by the arrows (or lack of them) in Figure 7 that are not consistent with the left to right route on which the classifier was trained.

How can we reconcile this result with the successful performance during our initial experiments? In our route following experiments we did not perform a full 360° scan as we did in constructing the map in Figure 7, but instead limited the scan to forward facing directions. If we construct a map similar to that in Figure 7 but for each point also provide a heading, then we can apply the navigation algorithm

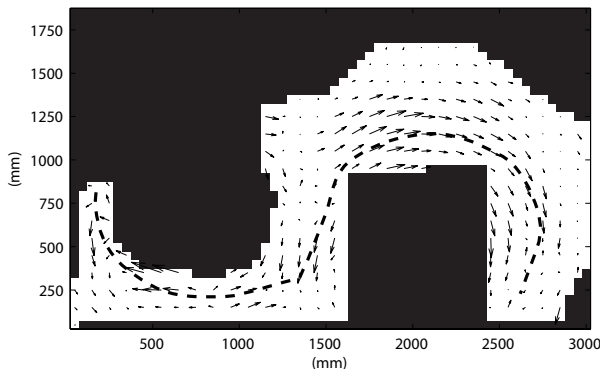


**Figure 7.** Map of predictions of the route direction based on the route classification algorithm for full  $360^\circ$  scans of the environment sampled across the permissible workspace. Confidence values are represented by the size of the arrows. The dotted line shows the successful route, going from left to right, followed in the first experiment. Black regions indicate areas of the workspace that were defined as out of bounds due to potential damage to the camera from collisions with objects.

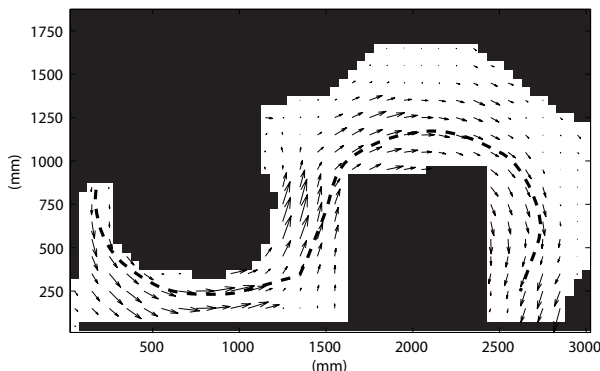


**Figure 8.** Map of predictions of the route direction based on the route classification algorithm for  $\pm 60^\circ$  scans of the environment centred on a heading parallel to the route heading of the nearest section of the original route. Confidence values are represented by the size of the arrows. The dotted line shows the successful route, going from left to right, followed in the first experiment. Black regions indicate areas of the workspace that were defined as out of bounds due to potential damage to the camera from collisions with objects.

exactly as it is used in the first experiment. We determine a heading for each point in space to be the route heading of the nearest section of the route. Figure 8 shows the preferred direction of movement together with confidence values again indicated by the length of the arrows. Note the area where the route failed in the 4th run of the first experiment has zero confidence indicated by the lack of an arrow which explains why the algorithm fails at this point. The algorithm appears to define a corridor along which successful route following is possible with confidence values decreasing the further away from the learned route that the agent strays.

## 4 Discussion

It is important to note that we are not primarily attempting an engineering solution to the problem of visual navigation. Instead we are trying to gain insight into how ants might learn and use visual information to guide their routes. This is an important point since it means that while successful route following using this approach is a minimum requirement, of potentially greater interest are the limitations and modes of failure that we observe. We have therefore not attempted to improve the overall navigational performance of the approach at the expense of the conceptual simplicity that it represents.

The idea that routes can be learnt using a set of Stimulus-Response (S-R) relations is not new. Equally, it has been observed by various authors that it is possible to orient rotationally by comparing views in different directions to a reference view, effectively resulting in a visual compass [17, 9]. However, combining aspects of these two approaches, as we have done, constitutes a novel approach. Firstly, by parameterising the S-R relationship using a boosted classifier, we not only provide a compact representation of the problem, we also obtain a more robust solution by being less reliant on determining an exact match between the learnt stimulus and the current view. Secondly, by using the classifier to determine view familiarity we are performing recognition rather than recall which is a fundamentally easier problem. In using familiarity rather than similarity to a particular reference view, we can go beyond a simple visual compass and instead construct a method for learning entire routes.

We have shown that it is possible to learn a non-trivial route through an environment using a simple view classification strategy based on positive and negative views collected during a single episode of learning. The route was designed to include high degrees of occlusion and variable depth structure such that a single snapshot taken at the route end-point could not underpin successful navigation through simple image matching. By considering the tight coupling of sensation and action that is present in ants and some robots we were able to reframe the problem of route navigation in terms of a search for simple directional views using a classifier that provides a compact way of storing the information required to recognise familiar views and crucially a measure of the expected uncertainty of the classification.

By embodying the view classifier on a physical platform and constraining the required spatial behaviour to routes, we were able to explore other areas for parsimony. The positive and negative views used by the classifier were collected by simulating a single sinuous path. Consequently, we observed that although spatial knowledge was fragile when the robot was placed at all points in the environment (Figure 7), as long as the agent has some context provided by the likely current direction of travel (Figure 8) the agent can recapitulate a learned route through a visually cluttered world and produce sensible headings from points off the original learned route. This provides an interesting example of where a simple interaction between a be-

havioural strategy and learnt information provides robust behaviour and without that interaction the agent would require a much more comprehensive survey of the environment. Interestingly, this type of interaction has been observed in ants where directional information from path integration has been shown to increase the precision of visual landmark use [7].

Our ultimate goal with this project is to understand likely and viable mechanisms used by insects for navigation. Therefore it is useful to summarise our framework with respect to some of the desirable properties of insect route behaviour: (i) Route knowledge should be procedural, i.e. an agent should be able to produce the correct behaviour for a given place independently of the prior sequence of visited places. By constraining vision and motion we produced a simple procedural mechanism for visually setting heading which is independent of the sequence of prior visited places. Although as noted above, some degree of hysteresis can be useful to compensate for a sparse set of positive and negative views acquired during training: An interaction which merits further study. (ii) Route knowledge should consist of a broad corridor of familiar places rather than a fragile narrow ridge and agents need to produce sensible behaviour when they are outside the route corridor. In our pilot-study we can satisfy this criteria as the estimates of heading produced from close off-route locations are sensible. Then as one moves further from the route the uncertainty in recalled headings increases which would be a useful signal to commence a systematic search for the route; a behaviour seen in ants when they are lost [8].

Our results suggest that it may be possible for ants and possibly other animals (including humans) to learn routes without the need for recall of the specific views encountered during learning. Instead, recognition together with the simple procedural rule of heading in the direction that appears most familiar may well provide sufficient information to allow successful navigation along routes.

## 5 Future Work

The ultimate test of our model of route acquisition in ants would require the comparison of the performance our approach with that of an ant in the same environment. There are obvious difficulties in achieving this, however as a first step towards this goal we intend to implement the algorithm on a mobile robot that can then be trained using images collected along an actual foraging route used by ants. It is hoped that by careful manipulation of prominent visual features along the route it will be possible to determine whether or not the approach provides a useful model of real behaviour.

In addition to this there are two main aspects of our approach that we are keen to explore further. These are: (i) How we represent an image. (ii) How we sample the environment to generate positive and negative examples for training a classifier.

In the current set of experiments we chose to represent views using the outputs of a set of Haar-like features. This choice was mainly motivated by the success of this approach when applied to a face detection task [13] where processing speed was a key factor. Since we are less concerned with how fast our code runs we intend to look at the result of employing a more comprehensive feature set consisting of Gabor filters at different positions, scales and orientations. By providing a pool of Gabor filters and learning routes in a variety of different ant-like environments we hope to determine whether there is any consistency in the filters that are selected by Adaboost, i.e. whether there is a general purpose set of filters that will work in a variety of different environments.

Next, we want to look at different ways of generating negative

training examples. Potentially any views that are not actually part of the route might be used to define negative instances for the purpose of training. We intend to explore how different, behaviourally plausible, sampling strategies effect the performance of the algorithm. For instance, in the current experiments we only provided negative views to the left and right of the current heading and found that this resulted in classification errors for views from the route facing backwards relative to the learned route direction. By including a wider range of negative views during learning it should be possible to improve the robustness of the approach.

## REFERENCES

- [1] B. A. Cartwright and T. S. Collett, 'Landmark learning in bees', *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, **151**(4), 521–543, (1983).
- [2] K. Cheng, A. Narendra, S. Sommer, and R. Wehner, 'Traveling in clutter: Navigation in the central Australian desert ant *Melophorus bagoti*', *Behavioural Processes*, **80**(3), 261–268, (2009).
- [3] M. Collett, T. S. Collett, S. Bisch, and R. Wehner, 'Local and global vectors in desert ant navigation', *Nature*, **394**, 269–272, (1998).
- [4] T. S. Collett, E. Dillmann, A. Giger, and R. Wehner, 'Visual landmarks and route following in desert ants', *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, **170**(4), 435–442, (1992).
- [5] V. Durier, P. Graham, and T. S. Collett, 'Snapshot memories and landmark guidance in wood ants', *Current Biology*, **13**(18), 1614–1618, (2003).
- [6] Y. Freund and R. E. Schapire, 'A decision-theoretic generalization of on-line learning and an application to boosting', *Computational Learning Theory: Eurocolt 95*, 23–37, (1995).
- [7] T. Fukushi and R. Wehner, 'Navigation in wood ants *Formica japonica*: context dependent use of landmarks', *Journal of Experimental Biology*, **207**(19), 3431–3439, (2004).
- [8] M. Kohler and R. Wehner, 'Idiosyncratic route-based memories in desert ants, *Melophorus bagoti*: How do they interact with path-integration vectors?', *Neurobiology of Learning and Memory*, **83**(1), 1–12, (2005).
- [9] F. Labrosse, 'The visual compass: Performance and limitations of an appearance-based method', *J. Field Robotics*, **23**(10), 913–941, (2006).
- [10] R. Rosengren and W. Fortelius, 'Ortstreue in foraging ants of the formica rufa group: hierarchy of orienting cues and long-term memory', *Insectes Sociaux*, **33**(3), 306–337, (1986).
- [11] L. Smith, A. Philippides, P. Graham, B. Baddeley, and P. Husbands, 'Linked local navigation for visual route guidance', *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, **15**(3), 257–271, (2007).
- [12] L. Smith, A. Philippides, P. Graham, and P. Husbands, 'Linked local visual navigation and robustness to motor noise and route displacement', in *SAB '08: Proceedings of the 10th international conference on Simulation of Adaptive Behavior*, pp. 179–188, Berlin, Heidelberg, (2008). Springer-Verlag.
- [13] P. Viola and M. Jones, 'Rapid object detection using a boosted cascade of simple features', *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, **1**, 511–518, (2001).
- [14] R. Wehner, 'Visual navigation in insects: coupling of egocentric and geocentric information', *Journal of Experimental Biology*, **199**(1), 129–40, (1996).
- [15] R. Wehner, 'The architecture of the desert ant's navigational toolkit (hymenoptera: Formicidae)', *Myrmecological News*, **12**, 85–96, (2009).
- [16] R. Wehner and F. Raber, 'Visual spatial memory in desert ants, *cataglyphis bicolor* (hymenoptera: Formicidae)', *Cellular and Molecular Life Sciences (CMLS)*, **35**(12), 1569–1571, (1979).
- [17] J. Zeil, M.I. Hofmann, and J.S. Chahl, 'Catchment areas of panoramic snapshots in outdoor scenes', *Journal of the Optical Society of America A*, **20**(3), 450–469, (2003).

# Constructing Emotions

## Epistemological groundings and applications in robotics for a synthetic approach to emotions

Luisa Damiano<sup>♦</sup> and Lola Cañamero<sup>♦</sup>

**Abstract.** Can the sciences of the artificial positively contribute to the scientific exploration of life and cognition? Can they actually improve the scientific knowledge of natural living and cognitive processes, from biological metabolism to reproduction, from conceptual mapping of the environment to logic reasoning, language, or even emotional expression? To these kinds of questions our article aims to answer in the affirmative. Its main object is the scientific emergent methodology often called the “synthetic approach”, which promotes the programmatic production of embodied and situated models of living and cognitive systems in order to explore aspects of life and cognition not accessible in natural systems and scenarios. The first part of this article presents and discusses the synthetic approach, and proposes an epistemological framework which promises to warrant genuine transmission of knowledge from the sciences of the artificial to the sciences of the natural. The second part of this article looks at the research applying the synthetic approach to the psychological study of emotional development. It shows how robotics, through the synthetic methodology, can develop a particular perspective on emotions, coherent with current psychological theories of emotional development and fitting well with the recent “cognitive extension” approach proposed by cognitive sciences and philosophy of mind.

### INTRODUCTION

“Understanding by building” [41] is the promise of the “synthetic approach”, currently presented as the methodology by which the sciences of the artificial can contribute to the scientific research on life and on cognition. Considerable work has been done by specialists of artificial sciences to define ways of “synthetically” modelling living and cognitive phenomena [7, 9, 16, 17, 33, 41, 42]. This renders even more important the question of how pertinent this modelling is for the scientific knowledge of life and cognition. What are the status and the value of the insights coming from the exploration of synthetic models of life and cognition? Is there a criterion to warrant a positive transmission of knowledge from the sciences of the artificial to the sciences of the natural? Can systems built with different materials, and endowed with different embodiments, be considered effective models of natural living and cognitive systems? Questions of this kind, if left unanswered, threaten the legitimacy of the synthetic approach, as well as its effective integration among the explorative practices accepted by the scientific community as a source of valuable insights. The main aim of this article is to generate answers for the aforementioned questions, by offering an epistemological solution to the issue of the legitimization of the synthetic approach. We will try to present this solution not only through speculative dissertation, but also through a concrete application taken from current scientific practice. In Section 1 we will introduce and describe an epistemological framework able to provide the synthetic approach with a general epistemological legitimacy, as well as with useful criteria to evaluate the

pertinence of its applications to the study of living and cognitive phenomena. In Section 2 we will present one concrete implementation of this approach, which, through the support of the epistemological framework previously proposed, explores a class of phenomena among the most problematic for science – emotions. The intent of this presentation is to show that the synthetic approach, when rigorously applied to emotional phenomena, can generate significant theoretical and methodological results. Indeed, it produces a rather innovative view of emotions, which heralds a generative two-way transmission of knowledge between epigenetic robotics and the psychology of emotional development.

### 1 AN EPISTEMOLOGICAL FRAMEWORK FOR AIIB AND AIICS<sup>1</sup>

#### Grounding the synthetic approach in constructivism

##### *a) The synthetic approach: the methodology of a creative science*

Increasingly, the sciences of the artificial claim they are able to go beyond a merely engineering approach. They manifest the ambition to structure a properly scientific approach, and to deal with the crucial questions at the basis of “sciences of the natural” such as biology, psychology and cognitive sciences. The intent is to offer to these sciences a privileged methodology to get to the hidden mechanisms of their objects: “understanding by building” [41]. In the current literature this emergent methodological principle promotes the programmatic production of embodied and situated models of living and cognitive systems in order to explore aspects of life and cognition usually not accessible in natural systems and scenarios. The procedural novelty of this approach, often called “constructive” or “synthetic” [7, 9, 16, 17, 33, 41, 42], is that it inverts the usual order between analysis of behaviour and construction of models. It requires the researcher firstly to “embed” the basic hypothesis on life and cognition in working artificial systems and then to examine the behaviours they produce [16]. It proposes a heuristics which opens to scientific investigation not only the question “how does it work?” but also “why this and not that?” [33] Besides this, it provides to science a new access to the simplicity of nature – better: a new concept of the simplicity of scientific explanation. The current versions of the synthetic approach share a common thesis on the genesis of complex behaviours, paradigmatically exposed in Herbert Simon’s “story of the ant” [46, 16, 40, 4]. This thesis, according to its most detailed elaborations, can be decomposed as follows:

<sup>♦</sup> STRI and School of Computer Science, University of Hertfordshire, College Lane, Hatfield, Herts AL10 9, U.K.

<sup>1</sup> We use AIIB as an abbreviation of Artificial Intelligence Inspired Biology, and AIICS as an abbreviation of Artificial Intelligence Inspired Cognitive Science.

- (1) Complex behaviours rest on the interaction of simple elements, but are not found in these elements taken separately.
- (2) Complex behaviours arise from the interplay between the interacting elements, the global system constituted by them and the environment with which the system interacts.
- (3) Complex behaviours tend to exceed the observer's possibilities of calculation and prevision, even when he built the system which manifests them.

This is clearly not a mechanistic perspective. The proponents of the synthetic approach express its "post-mechanistic" character through the notion of "emergence" – a sort of emblem of the arising "post-classical" science [14].<sup>2</sup> This notion not only disobeys the old mechanistic postulate "the whole is the sum of its parts". It also refuses the mechanistic heuristic pretension to a science able to know past, present and future behaviours of the objects it studies – Laplace's demon "omniscient science" [3, 43, 47]. And it is through the dismissal of these two traditional schemes of scientific rationality that the concept of emergence promises to the synthetic approach effective and simple models of life and cognition. Hypothesis (1) gives to science the possibility to plan the construction of artificial models of living and cognitive systems as the organisation of elemental components in integrated dynamical units – that is, as the manipulation of relations between elemental components.<sup>3</sup> Hypothesis (2) allows researchers to ideate and design for these systems quite simple internal mechanisms, as it grounds the complexity of living and cognitive behaviours not in the systems, but in the interplay between them and the environment. Hypothesis (3) authorizes scientists to consider these artificial models of living and cognitive systems as actual generators of knowledge, since, according to this hypothesis, once these models are situated in an environment, they can express behaviours unexpected by their constructor, and give him new insights and feedbacks about the thesis on life and cognition that they embed.

<sup>2</sup> The divergence between emergentism and mechanism we invoke in the text refers to mechanism as the philosophical (ontological and epistemological) view at the basis of "modern" or "Newtonian" scientific tradition. To simplify, this view can be related to two main elements: (a) the idea of the natural universe as a deterministic succession of qualitatively homogeneous states (a "clockwork universe" wherein everything can be deduced from a certain starting point), and (b) the idea of the observer as a cognitive agent endowed with an external and neutral point of view (an "absolute spectator of nature"). These elements (which, for the sake of brevity, are over-schematized here) ground the scientific heuristics which defines what some epistemological literature, as well as this article, calls "classical science" [3, 43, 47]. This heuristics corresponds to the ideal of elaborating a description of nature endowed with five primary characteristics: determinism, objectivity, simplicity, homogeneity, completeness, well-described in [43]. With the adjective "post-classical" or "non-classical" we refer to every form of science which departs from this heuristics as well as from the view of nature and of science on which it rests. In this article emergentism is defined post-classical in this sense, and not in the sense that it implies the refusal of "mechanistic explanations", i.e. explanations which rely on the identification of mechanisms and avoid the introduction of meta-physical forces (Sect. 2.c).

<sup>3</sup> According to the emergentist point of view, organisational interrelations between elemental components can inhibit some of their individual behaviours and generate collective behaviours which are not expressed in the elements taken separately [14].

But this theoretical structure, although offering to the synthetic approach the advantage of a simple explanation, does not spare it the disadvantages of a controversial explanation of living and cognitive processes. This emergentist framework, relying on one of the most problematic and debated notions of contemporary science, is not a good candidate for solving doubts about the contribution of the science of the artificial to the scientific understanding of natural phenomena.

Even if we could obtain full agreement on the idea that the creativity of nature can be conceived in emergentist terms, many unanswered critical questions would remain. Can the artificial construction of natural phenomena actually reproduce the creative action of nature? Can it actually accede to the natural mechanisms creating life and cognition? Can it genuinely penetrate and reconstruct – and not roughly imitate – complex processes such as biological metabolism and reproduction, or conceptual mapping of the environment, linguistic expression, emotion? In short: in what sense is a significant transmission of knowledge from the sciences of the artificial to the sciences of the natural possible?

History of science shows that this kind of transmissions has taken place since the birth of the sciences of the artificial, and with remarkable results. Two of the most significant examples are the application in biology of the cybernetic concept of feedback, fundamental for the scientific modelling of cellular metabolism and biological stability [14], and the transfer of the scheme of the digital computer from the engineered production of artefacts to the scientific description of natural cognition [4]. This transfer of knowledge has the merit to have given rise to the classic or computationalist form of cognitive sciences, and to have created the paradigm which oriented their development until recent times – actually it still does, albeit partially.<sup>4</sup> Besides, philosophy of science has lately tended to acknowledge and enhance the positive contribution of the sciences of the artificial to the scientific understanding of nature [1, 44]. For example, some of the latter's new trends, such as philosophy of scientific instruments and philosophy of technology, are developing the thesis that the stable production of phenomena through artefacts could allow one to overcome the problem of induction [20]. Moreover, current scientific literature is increasingly engaged in elaborating procedures directed to implement and make rigorous the application of the sciences of the artificial to research on natural phenomena. But history, philosophy and methodology leave open the issue of the epistemological legitimacy of this emerging way of doing science.

The synthetic approach does not simply subordinate analysis – the classically privileged direction of scientific investigation – to synthesis – indeed, a synthesis which intends to diverge from the mechanistic "sum of parts". This approach challenges the "representationalist" epistemology typical of modern science.<sup>5</sup> It

<sup>4</sup> Since the 1980s, the computationalist paradigm is strongly criticised, mainly for its poor biological plausibility [38]. In the 1990s, from the criticisms to computationalism arose a new candidate to the guide of cognitive sciences, namely, the paradigm of "embodiment" [12]. This is supported by trends of research which intend to overcome the Cartesian dichotomy inherited by computationalism – that is, they intend to "re-embodiment the mind" [50]. For the most part, current research uses descriptive solutions which hybrid the two paradigms [12].

<sup>5</sup> Representationalism is the way of conceiving knowledge typical of classical science as defined before (Footnote 2). It grounds the ideal of

promotes a way of describing nature which diverges from the old-fashioned ideal of an exhaustive reflection of the natural world free from subjective alterations. “Understanding by building” proposes to practice science as a deliberate act of construction, exercised on systems which would not exist without this act. It demands that researchers think and implement science as a form of knowledge which actively creates, and does not passively reflect, the phenomena explored. It stimulates the elaboration of new epistemological groundings for the scientific enterprise – epistemological principles able to generate answers for the critical questions which threaten the legitimacy of the new creative approach to the modelling of nature.<sup>6</sup>

### ***b) Scientific constructivism: an epistemological framework for the synthetic approach***

The hypothesis we propose in this article is that some epistemological groundings appropriate for the new creative science already exist. They have been elaborated by a heterodox branch of twentieth-century science, made up of the first groups of research to introduce and develop the notions of autonomy and self-organisation in biology and cognitive sciences. They are a few mutually independent groups, among which some were dedicated to naturalistic inquiries (i.e. organicistic embryology groups, such as, the Cambridge group and the Brussels group, and the thermodynamics of dissipative structures group, that is, the School of Brussels founded by Ilya Prigogine), and some to cybernetic research (e.g. the Biological Computer Laboratory group founded by Heinz von Foerster<sup>7</sup> and the French neoconnectionist group of Henri Atlan). Usually the main contribution attributed to them is the new vision of nature generated by their exploration of the “endogenously controlled organised systems” – i.e. the systems that they labelled as “autonomous” or “self-organising”. But these first explorers offered to contemporary science another relevant contribution, often neglected. It consists in a new scientific epistemology, aimed at underpinning a tradition of research which conceives and practices scientific knowledge not as the “representation of nature”, but as its “construction” (i.e. a “co-construction” due to the interaction of the observer with the reality he explores).<sup>8</sup>

---

the objectivity of scientific knowledge as independence from its subject or subjective aspects [3, 43, 45].

<sup>6</sup> The underlying hypothesis is that objectivist representationalism (i.e., to put it roughly, the doctrine according to which the knowing subject, via his sensory apparatus, can dispose of internal representations of external and pre-determined objects) is not the best gnoseological option for the synthetic approach. This epistemological doctrine tends to describe human knowledge as the passive internalisation of the external world and, as a consequence, tends to see any active role of determination exercised by the subject on his objects of knowledge as a source of alteration. This leads us to propose for the grounding of the synthetic approach a constructivist epistemology, as it sees the active role of the subject in the process of cognition not as a source of subjective contamination of his knowledge of reality, but as a necessary ingredient of it, which, in certain conditions, can produce pertinent knowledge. Our option, which orients the development of this article, does not intend to imply that all the proponents of the synthetic approach are or must be constructivist, nor it deny that often they adhere to a representationalist epistemology.

<sup>7</sup> As well-known, this laboratory hosted researchers such as Gordon Pask, Georg Zopf and Ross Ashby.

<sup>8</sup> The contents of the present sub-section of the article are dealt with in detail in [14].

Indeed, the notion of self-organisation – a notion which is closely linked to that of emergence – is strongly heterodox. It is not merely the core concept of a post-mechanistic scenario in which natural evolution manifests a property denied by the Newtonian tradition: the creativity required to overcome the physico-chemical level of reality and give rise to “qualitatively different” or “emergent” levels – life, cognition, man, human scientific and technological creations. The notion of self-organisation, by supporting this view of natural evolution, is also at the heart of the pioneers’ transition to a new scientific epistemology. As they adopted this emergentist theoretical background, the early researchers on self-organisation rejected the classical image of the observer as an “absolute spectator” of nature, without localization and perspective. They described the scientific observer as belonging to the natural world and approaching nature from within – from a “limited and situated point of view”. According to their evolutionary conception, the observer is an embodied material system belonging to the class of the natural self-organising systems. With every other system of this kind, he can experience external events insofar as they destabilize its internal dynamics and to the extent he can attribute them operational meanings through self-regulation. His knowledge is made of interpretations determined not only by its physico-biological organisation, but also by the specific characteristics of the socio-cultural systems to which he belongs. He cannot develop categories able to generate neutral representations of reality. His theoretical categories play a selective and constructive role, and limit his descriptive domain to those aspects of nature they can define – “build” – as determined objects of research. These categories offer to the scientist not an objective knowledge in the classical sense, but a “pertinent” knowledge, that is, knowledge able to warrant him effective operationality in the domain to which they can be applied.<sup>9</sup>

This transition from representationalism to constructivism was not only due to speculation. For the pioneers it was primarily due to the need of providing an effective description of the systems they studied.

Actually, the “internal determination” typical of autonomous systems – the fact that, by self-regulation, they can break the causality of external variations and determine their internal variations – corresponds for science to a limit of intelligibility. This kind of “endogenous control” implies that autonomous systems can resist not only environmental pressures, but also scientific knowing actions. The observation and the manipulation of this class of systems go with an uncertainty which deprive of any plausibility the classical scientific heuristics. With respect to autonomous systems science cannot satisfy the request of knowing past, present and future behaviours of the objects it explores – “capturing their dynamical law”. In other words: with regard to these systems science cannot produce a standard classic description [3, 43].<sup>10</sup> Their unpredictable transformations oblige the observer to permanently (re)negotiate with its objects the

---

<sup>9</sup> The primary references are [3, 21, 25, 39, 43, 47].

<sup>10</sup> As pointed out by some of the pioneers [3, 21, 39, 43], to describe autonomous systems requires to depart from the classical ideal of the scientific “representation” of nature (Footnote 2) – i.e. to produce a description which is not deterministic, not simple (no fundamental levels of observation result available; to accede to all the observable content of the system different levels of observation are needed), not homogeneous (more than one system of description is needed), not complete (depending on the descriptive system used, different observables are available), not objective in the classical sense (dependant on the subjective levels of observation and systems of description).

categories and the pertinent observables of their scientific characterisation. When they are applied to autonomous systems, scientific categories can only produce partial and revocable descriptive solutions. They express a kind of functioning which forces the observer to definitively decline the traditional idea of panoramic inspections of nature, and practice scientific research as the construction of plausible referents for entities which in themselves are not accessible. Concretely, this practice corresponds to a post-classical procedure of characterisation: to structure and coordinate an increasing multiplicity of theoretical levels of description, each able to define as a definite object of research a different aspect of the system explored. In short: to find and to study carefully an autonomous system's resistances to the application of a model; to exploit this study for the development of new theoretical points of view on the system; to move from one to another of these point of views, tracking the system's manifestations – its transformations. The pioneers thought this heuristics as a style of knowledge which requires the researcher to learn from nature how to build it as a set of definite objects of research.<sup>11</sup> And they conceived the artificial construction as a useful complement of this theoretical construction of nature.<sup>12</sup>

The epistemological thesis at the basis of this new heuristics radicalises the outcomes of the “crisis” which in the last century weakened the foundations of classical science. According to the pioneers' view, the action of determination exercised by the observer on the objects he explores is not the contingent alteration of a reality which is neutrally accessible. Indeed, it is a positive and constitutive ingredient of the scientific description of nature: an essential creative activity, without which reality cannot manifest the form of defined objects of research.

On this thesis the first explorers of self-organisation built a new and complex epistemological framework for science, which deeply transforms the structure of the classical one. At a conceptual level, it integrates the traditional epistemological dichotomies, but modifies their internal configuration [3, 43, 47]. It transforms their opposite terms in complementary terms, and allows science to adopt a kind of intelligibility in which *discovery* can be *invention*, scientific *facts* can be *artefacts*, *objective evidence* can converge with *subjective construction*, and the old-fashioned “spontaneous manifestations of nature” can be found in its theoretical and artificial scientific *reifications*.

This is the epistemological framework we propose as pertinent for the grounding of the current synthetic approach. In the remainder of this section we will focus on two of its basic epistemological principles, introduced to support a modelling devoted to “generate” the natural behaviours explored, and not to “formally represent their law”. We will propose these principles not in their early versions, but in the most elaborated one. It was introduced by the School of Santiago founded by Humberto Maturana and Francisco Varela – i.e. the autopoietic biology group, which, in a sense, can be considered a descendant of the

pioneer groups [33-34].<sup>13</sup> We believe that these two principles could offer to the synthetic approach not only a general epistemological legitimacy, but also some criteria useful to evaluate the pertinence of its applications to the study of life and cognition.

### c) Two epistemological principles for AIIB and AIICS

I. The first principle proposes a constructivist definition of scientific explanation, paradigmatically expressed in many places of Maturana and Varela's literature.

“If you want to explain lightning, you must provide a mechanism that generates it.” (H. Maturana)<sup>14</sup>

The epistemological content of this principle can be conceived as an *operational* concept of scientific explanation, according to which explaining a phenomenon amounts to proposing a mechanism able to produce it. The pioneers' constructivist style of knowledge finds here an evolved expression. This postulate does more than presenting the equation in which is grounded the first explorers' creative way of doing science – to know scientifically is to “build” (“construct”, “invent”, “fabricate”). It introduces a specific “post-classical” procedure of description, which integrates and enhances the heuristic attitude of the early research on autonomy. This descriptive solution can be seen as particularly appropriate to deal with autonomous systems, since it cannot be affected by their unpredictability. Requiring models able not to predict, but to generate natural processes, the principle locates the focus of the scientific research not anymore on actual, but on possible natural behaviours. It discards the classical demand of predicting and controlling nature, and proposes a constructive explanation which is destined to be revocable. In the School of Santiago's literature, it is associated to the heuristic imperative of evaluating the progressive character of the explanation, that is, establishing if the mechanism that the explanation proposes is able to produce other phenomena belonging to the same domain. If the explanation does not result progressive, it has to be substituted with a more generative one; otherwise, it has to be re-tested. This re-proposes through procedural terms the main epistemological imperative of the constructivist logic of description: do not impose to reality a structure which belongs to human creativity.

“If I'd like to provide a scientific explanation of cognition, I must provide a mechanism that generates an appropriate (animal or human) behaviour, and other behaviours which are susceptible to be observed in the same domain.” (H. Maturana)<sup>15</sup>

This operational concept of scientific explanation (with its new emphasis on *possible* rather than *actual* behaviours; on *generation* rather than *prevision* of phenomena; on *construction* rather than *representation* of objects) produces the epistemological shift that we consider appropriate for the

<sup>11</sup> See the references given at Footnotes 9 and 10. There is not enough room here to describe in detail the constructivist conception of science. A good description, focused on the substitution of the classical idea of objectivity with that of pertinence or viability, can be found in [24, 45].

<sup>12</sup> This is particularly true for the cybernetic pioneer groups, which, in their scientific explorations, tried to systematically couple artificial and theoretical construction. In this sense, they can be considered as the founders and pioneer implementers of the current synthetic approach.

<sup>13</sup> The School of Santiago (in particular Maturana) denied or minimized the influence of the pioneer groups' productions on its theory of life, although this influence is difficult to underestimate at many levels. Autopoietic biology exhibits very strong convergences, both theoretical and heuristic, with the pioneers' production, and the contacts of its authors with the pioneers (e.g. Heinz von Foerster) and their literature are well-known, as well as acknowledged by Maturana and Varela.

<sup>14</sup> See [34], p. 80

<sup>15</sup> See [34], p. 81

legitimization of the synthetic approach.<sup>16</sup> Its pertinence for the epistemological grounding of this methodology becomes flagrant if we consider its autopoietic application to the description of life.

On the basis of this epistemological principle, autopoietic biology formulated a procedurally new definition of life, which, instead of listing the main features of living systems, provides the theoretical draw of a dynamical mechanism able to produce their phenomenology [35]. Maturana and Varela called their definition of life “synthetic”, to distinguish it from the traditional “analytic” definitions presenting detailed lists of properties. And, as it is manifest in their work, they grounded this synthetic definition in an emergentist logic, made up of the same thesis nowadays adopted by the supporters of the synthetic approach: (1) although living phenomenology rests on physico-chemical components, it cannot be found in them; (2) it emerges from the interaction of these components, the unitary system they compose and its *ambience*; (3) resulting from this interplay, living phenomenology exceed science’s power of calculation and prevision [14, 32].

Affinity with the current synthetic approach is so strong that it authorizes us to hypothesize a direct link: autopoietic biology could be a historical source of methodological inspiration for the current synthetic approach, although its proponents usually do not refer to it as such. One of the rare case in which autopoietic biology is quoted in their literature is provided by the works of a nascent trend of synthetic biology (SB) called “chemical autopoiesis” [32], in which autopoietic biology figures explicitly as a theoretical source, and implicitly as a methodological one. This emerging branch is strongly engaged in overcoming the dominant bioengineer approach, and tries to answer to crucial scientific questions on life and cognition through the chemical implementation of the autopoietic definition of the minimal living system [33].

For these kinds of attempts, as to every synthetic approach expressed in SB or AI, the autopoietic operational concept of the scientific explanation offers a general epistemological legitimization, but with an imperative clause. To produce genuine insights, the artificial models of living and cognitive systems must display the same organisation of natural living and cognitive systems (see below).

II. The second autopoietic principle we consider pertinent for the epistemological groundings of the synthetic approach is indeed a theoretical postulate; but, as we will propose, in the context of the sciences of the artificial it assumes a significant epistemological value.

The basic content of this theoretical principle is the distinction between two notions, i.e. *organization* and *structure*. Simplifying the original autopoietic formulation, we can put it as follows:

- the *organisation* of a living system is its relational frame, that is, the network of functional relations which define the system as a unity of components;

- the *structure* of a living system is its materialization, given by the actual components and their interconnections [35].

This distinction is not a theoretical novelty introduced by Maturana and Varela. It was progressively elaborated by the

early researchers on living and cognitive autonomy. A first complete formulation can be attributed to Jean Piaget [40], who realized an integrative elaboration of the pioneers’ studies which, as showed elsewhere [14], seems to have strongly influenced Maturana and Varela’s production. Piaget proposed this conceptual distinction as the theoretical key to the comprehension of biological systems as dynamical, since it corresponds to the distinction between the invariant and the variant aspects of their dynamics. As he remarked, living systems can be considered dynamical systems endowed with a peculiarity: all their elementary components permanently change, while systems, as relational unities of components, remain. This, as Piaget pointed out, can be affirmed at both the ontogenetic and the phylogenetic levels. The relational unity is what remains unchanged not only in the permanent flux of physico-chemical component typical of biological organisms, but also during the ontogenetic transformations which can make a living system unrecognisable from one observation to the next. Moreover, this relational unity is transmitted through reproduction and remains unchanged generation after generation. Indeed, this relational unity is the invariant of the biological dynamics and therefore the lowest common denominator of living systems. To distinguish this invariant relational frame from the changeable materializations of living systems, and to determine its configuration, amounts to isolating an element which defines the class of dynamical systems belonging to the biological domain.

Here lies the relevance of the distinction between organization and structure, which is at least double. Firstly, this distinction opens the possibility of giving an operational explication of life, that is, the chance, exploited by autopoietic biology, to define a mechanism able to generate the living dynamics.<sup>17</sup> Secondly, this distinction generates significant implications both for theoretical biology and the epistemology of the sciences of the artificial. Synthetically: (a) In principle the materialization of living systems can be manifold;<sup>18</sup> (b) An artificial model of living systems, which is built with different materials and therefore endowed with a different embodiment, can be considered belonging to the class of living systems if it shares their organization; (c) To obtain pertinent insights for the study of life, the sciences of the artificial must produce models of living systems which display the same organisation as natural living systems.<sup>19</sup> Given the autopoietic equation between life and

<sup>17</sup> As Maturana and Varela pointed out [35], this must be a mechanism able to produce the living organizational invariance through permanent structural variation.

<sup>18</sup> The only clause is that the materialization of a system must allow its organizational invariance [2].

<sup>19</sup> This is the pertinence of insights which come from the exploration of systems belonging to the same class of living systems. The issue of establishing if the sciences of the artificial could produce systems endowed with the organisation of living and cognitive systems deserves a specific treatment. Here we would like to merely suggest that this can be done with different degrees of approximation, and according to different theoretical choices, in order to obtain insights whose pertinence is relative to these choices and has many possible degrees. If research wants to adopt autopoietic biology not only as a methodological, but also as a theoretical framework, then it has to take into account that autopoiesis defines in detail only the organisation of the minimal living cell. Indeed, Maturana and Varela’s theory describes very vaguely the organisation of superior organisms. For organisms endowed with the nervous system, autopoiesis provides only the constructivist draw of a circular organisation which connects “sensorium” and “motorium” [35].

<sup>16</sup> As pointed out by some of the pioneers and of their descendants, this notion of explanation is not an absolute novelty, but has some philosophical precursors, such as Thomas Hobbes and Giambattista Vico [23].

cognition, (a), (b) and (c) has to be considered valid for cognitive systems too.

III. The two autopoietic principles presented above produce two criteria which can be useful to evaluate the applications of the synthetic approach to the study of life and cognition. To sum up: in order to produce pertinent insights, the artificial models of living and cognitive systems must

1. have the same organisation of living and cognitive systems, although they can have a different structure;
2. embed a mechanism able to produce the living and cognitive phenomena they intend to explain.

These criteria, with the epistemological framework proposed, guide our group's application of the synthetic approach to the study of emotions – phenomena lately reintegrated within the proper phenomenology of cognition.<sup>20</sup> In the following section we will try to point out the theoretical and methodological contributions of this approach to the understanding of emotions.

## 2 MODELLING EMERGENT MINDS' EMOTIONS

### The generative loop between robotics and psychology

It is worth noticing that the application of the synthetic approach to emotional phenomena is not theoretically neutral. It requires researchers to develop and implement a very specific conception of emotions – indeed, a promising one. Besides being quite original in comparison to the classic philosophical and scientific views of emotions, this theoretical perspective fits well with some of the most recent developments in cognitive science and neurophysiology. Moreover, it shows a strong affinity with some new theoretical trends of developmental psychology, to the extent that a science of the artificial like robotics seems able to provide to psychological theories on emotional development a pertinent and fruitful experimental test-bed.

In robotics the rigorous adoption of the synthetic methodology amounts to departing from those AI approaches that exploit symbolic architectures (i.e. architectures grounded in rule-based symbol systems) in order to generate emotions by specific modules devoted to their computation, and to make emotions perceived by introspection [7]. In this case the result is a computationalist view of emotions which re-propose a conceptual element typical of classical western philosophical view. According to this approach, emotions are conceived and built as private events, which arise within the intra-individual space as internal states resulting from internal evaluations. They are basically an individual creation, accessible primarily to the subject of emotional experience and then to other subjects, through the agent expression.

This aspect of the classical theory of emotions is rejected by the synthetic methodology, whose emergentist framework requires developing different approaches to the artificial generation of emotions. More specifically, it requires researchers to construct emotions not anymore as integral part of the agent architecture, but as arising from the interactions of elements which in themselves do not constitute a internal emotional machinery. Indeed, according to the emergentist background seen above, emotional behaviours must be modelled as emerging from the interactions of elements which do not belong only to the agent, but also the environment. This approach, developed by both symbolic and embodied tradition of AI [7, 42], finds a radical

application in Valentino Braitenberg's fundamental text *Vehicles* [6]. In Braitenberg's drawings agents display architectures which contain only sensorimotor correlations and produce emotional behaviours through the interaction with some perceivable aspects of their environment. As pointed out elsewhere [7], this design of emotions has the merit of expressing very clearly the main characteristics of the synthetic approach, but has also the default of bringing them to their very limit, and risks resulting unproductive. Braitenberg draws architectures deprived of any elements to which the production of emotions, as known in natural systems, can be related. He models emotions as pure epiphenomena "in the eye of the observer", and therefore misses the possibility of studying the role played by emotional processes in natural and artificial agents' interactions with the environment. More productive applications of the synthetic approach tend instead to explore how emotional behaviours arise from the interaction of "underlying mechanisms" inspired by the scientific knowledge of natural emotional processes. This procedure does not implement in robotic architectures the notions used to describe emotions, but those used to describe lower level processes, and it tests these implementations in agents interacting with the environment.<sup>21</sup>

In both these versions, the synthetic approach supports a view of emotions which discards the elements belonging to their classical philosophical conception. It declines the idea of emotions as private events arising within an individual and separable organisation. It locates emotions' generation not in the agents in themselves, but in their relation of coupling with the environment. It finds the pertinent unit for the study and the modelling of emotional processes in the whole system-environment unit – not an individual, but a *relational* unit, and primarily an *inter-individual* unit.

This idea, which in general is not explicitly formulated, connects the synthetic approach to emotional phenomena with a recent trend in the theory of emotions and, again, to the early research on self-organisation. Indeed, this rising trend is developed by researchers in different fields of natural and human sciences who share a strong interest for the theory of self-organisation [19, 31]. They inherited from the early sciences of autonomy the thesis that self-organising systems, such as living and cognitive systems, can experience dynamics of correlation which transform them in sub-units of higher level adaptive units. This idea grounded the pioneers' view of inter-subjectivity as the correlation of individuals in inter-individual units. To sum up: the mutual interaction between cognitive autonomous systems, for example human individuals, produces the co-dependence of their somatic and neural networks' dynamics, and couples their self-determined cognitive and affective behaviours at many levels.

Concepts such as Piaget's "inter-individual regulation" [40], autopoietic "behavioral coupling" [35], first Paskian and then Varelian "conversational unit" [39, 49] were introduced to express this idea. They can be seen as conceptual ancestors of the arising self-organisational theories of emotions and emotional development. The hypothesis that these theories develop is indeed one of the main implication of the self-organisational conception of inter-subjectivity: "emotions are social and [...] the body proper to emotions is the social body" [19]. In this picture emotions are not individual, but common –

<sup>20</sup> This re-integration is due to the recent attempts to "re-introduce the body in cognitive science" [50, 38].

<sup>21</sup> Typically this kind of design uses homeostatic control and neural networks modelled at different levels of abstraction [7].

“common works” [19]. They are aspects of an inter-individual process of coordination of intentions and actions in which social individual beings are involved from their birth and which, by affecting the most basic levels of their bodily organisation, deeply shapes their interaction with the environment. This thesis contrasts sharply with the computationalist view of emotions. For example, it excludes the possibility that the basic operation by which we accede to others’ emotions can be found in a rational evaluation relying on the analysis of their expressions. It affirms that, before any rational calculation, we experience others’ emotional processes, since we are essentially involved in these dynamics at the very basic levels of our biological and cognitive organisation [15]. There is no room here for the detailed explanation this view deserves. But it is worth noting that this kind of theory of emotions, besides converging with that implicit in the synthetic approach, is increasingly accepted and developed. It seems to express a basic element of inter-subjectivity which has lately been detected and explored by different branches of cognitive sciences.

Since the late Nineties, neuroscience has been studying the neural mechanisms underlying inter-subjective emotional processes – Vittorio Gallese, one of the discoverers of “mirror neurons”, calls them “mirror mechanisms” [22].<sup>22</sup> The evidence pointed out by these studies is that the observation of one agent’s emotional expression activates in the observer the neural pattern underlying this expression. This implies that in recurrent inter-subjective interactions there is such a co-dependence of the “respective” emotional dynamics that it requires us not to think of them as individual, but as intrinsically inter-individual processes.

This idea of an “inter-individual emotionality” converges with some of the last and most interesting theoretical advancements in cognitive sciences and philosophy of mind. They consist in new designs of the “embodied mind”, i.e. the cognitive architecture by which researchers try to overcome the Cartesian dichotomy inherited by computational cognitive science. The current proponents of the *embodied mind* are engaged in a theoretical operation that philosophy of mind calls “cognitive extension” [51]. It is directed to redefine the boundaries of mind in line with this insight: if mind is the area where cognitive processes take place, then it cannot be relegated within the intra-individual space. According to the most widespread version of this thesis – a version labelled with the concept of “the extended mind” – the region of the mind overcomes the limits of “skull and skin” to include pertinent elements of the environment, i.e. elements which are external to individual body and brain, and which are necessary to the accomplishment of the agent’s cognitive processes [13]. But there is also a more interesting version of the same thesis, which transforms the idea of the “cognitive extension”. Its proponents give to the problem of “the embodiment of mind” not an anatomic, but a dynamic solution. They ground the mind not in the nervous system’s anatomy, but in the regulative dynamics by which it links the body to the environment. Their “extension” of mind does not overcome spatially “skull and skin” [13]; instead, it re-defines the architecture of the cognitive mind. It describes the mind as the structure of coupling which inter-connects the dynamics of nervous system, body, environment and other organisms – other

selves. In other words, it designs the mind not as a spatial – “extended” [13]– object, but as the dynamical co-determination which couples self and others’ somatic and neural networks between them and to their environment. In this picture mind is essentially this: the emerging unit of self, others and environment’s co-determination, which produces the cognition of the others and of the world. This view has been developed rigorously by a current trend of embodied cognitive sciences called “radical embodiment” [12, 14], which explicitly includes in its genealogy the early research on autonomy and autopoietic biology. Its concept of mind, often called “radically embodied mind”, expresses with the tools of dynamical systems theory the theoretical content that the notion of “extended mind” tries to express too: cognitive processes – emotional processes included – do not belong to an individual, but to a *relational* unit – for social beings primarily the “self-other” *inter-individual* unit.

This emergentist and inter-individual concept of mind offers a theoretical expression not only to the synthetic approach to the study of emotions, but also to a new trend of current psychology of emotional development [36]. More specifically, this notion offers to them a conceptual link, which nowadays seems able to connect robotics and psychology in a circuit of exchanges susceptible to improve the scientific understanding of emotional development.

The specificity of this trend of psychology of emotional development is that it elaborates the idea of the inter-individual character of these developmental processes. A remarkable example is offered by the theory of attachment [11], in which the child and the caretaker(s) are not presented as independently defined individuals who undertake recurrent interactions. Instead, they are described as the polarities of a “dyad”: an inter-individual system whose components are involved in a dynamics of co-determination which shapes the child’s way of cognitively and affectively interact with his (social) environment – and reshapes the caregiver’s way. This inter-individual view of emotional development is typical not only of the original version of the theory of attachment [5], but also of its critical and revised versions [30]. Some of these latter are formulated in self-organisational terms, [48] and offer to synthetic approach and developmental psychology, besides strong theoretical affinities, a partially shared scientific genealogy.

This convergence between the synthetic approach to the study of emotions and psychology of emotional development is the premise of the collaboration of our epigenetic robotics’ group with psychologists. The main goal is structuring an interdisciplinary exploration of emotional development and, in particular, of the mechanisms underlying the creation and the evolution of attachment bonds. This co-exploration rests on the joint, interdisciplinary design of robotic experimental scenarios, which are inspired by psychological insights and aim to produce feedback useful for psychology of emotional development and attachment.

Concretely, this collaboration is producing a series of studies focused on attachment bonds linking a “baby” robot and one (or more) human caregiver(s). Several crucial aspects of attachment are under inquiry: the development of different attachment profiles, the influence of these different profiles on exploratory behaviours, the role of attachment bonds in the development of sensorimotor associations, the development of attachment bonds in presence of multiple caregivers... [10, 26-29] A good premise for the pertinence of these robotics experiments for psychology is in our attempt to respect the two criteria presented above: (1)

<sup>22</sup> The notion of “mirror mechanisms” is introduced by Gallese to avoid the vague use of the notion of “mirror neurons” criticised in the current debate [18].

to give to the robots, as much as possible, a cognitive organisation which reproduces that of the natural cognitive systems they intend to model; (2) to make the robots generate, in interaction with humans, the phenomenology described as typical of attachment in specialists' literature. We are trying to satisfy criterion (1) through the implementation of artificial nervous systems which are not bio-mimetic, but are based on a perception-action organisation.<sup>23</sup> In these architectures, nothing plays the function of a internal emotional machinery – there are only mechanisms which allow the robots to modify their behaviours according to the interaction with the human user. These mechanisms, described in detail elsewhere, can satisfy criterion (2), since they are able to produce the fundamental aspects of attachment, such as the active research of the presence and the attention of the caregiver, distress in situation of separation, changes in attachment profile and exploratory behaviours in function of the behaviour of the caregiver etc... What this kind of architectures tries to implement is the thesis of the inter-individual character of emotional and attachment processes, which are not built as integral part of the robots, but as the product of their interactions with the users within a dynamical environment.

We believe the results we are obtaining are useful for both robotics and psychology, and moreover for the interdisciplinary unit they constitute. For robotics they consist primarily in significant steps towards the design of robots capable of an emotional development which make them adaptable to their specific user(s) and therefore more apt to interactive roles, such as the role of companions. With respect to psychology, these studies seem very useful for psychologists to evaluate and improve its theories of emotional development by testing what usually cannot be tested because of ethical constraints (e.g. the production of negative attachment bonds) or limits of intelligibility (e.g. the mechanisms underlying different attachment bonds). Finally, this collaboration heralds the development of a two-way interdisciplinary enterprise able to improve the scientific understanding of emotional development and, more generally, human cognition and mind. These studies allow us not only to test and possibly improve (or reformulate) the idea of inter-individual nature of emotional and cognitive processes, but also to offer our implementation and testing of this notion to new fields of research, such as the psychology of human-machine (human-robot) interaction.

We believe that this kind of cooperation – even if currently it is only at a starting point – can give a picture of what a science integrative of the synthetic approach could be: an interplay between sciences of artificial and sciences of the natural which makes productive the constructivist equation between *scientific facts* and *artefacts*.

## CONCLUSION

In this article we have dealt with the issue of warranting a positive transmission of knowledge from the science of the artificial to the natural sciences of life and cognition. More specifically, we have tried to show how it is possible to warrant the pertinence of the “synthetic methodology” – “understanding by building” – for the scientific research on living and cognitive phenomena. Our contribution has not been only speculative, but has integrated an epistemological proposal for the legitimization of the synthetic approach with a concrete example of its

application in the scientific practice. In Section 1 we proposed for the epistemological groundings of the synthetic approach the constructivist epistemological framework defined by the early research on self-organisation. In particular, we have presented two principles able to offer an epistemological basis to this approach. They are principles of intelligibility extracted from autopoietic biology, which respectively define: (a) an operational concept of scientific explanation and (b) the conceptual and methodological distinction between organisation and structure. From these two principles we have derived two criteria useful to evaluate the applications of the synthetic approach. In Section 2 we presented a concrete application of this approach, which, through the support of the proposed epistemological framework, explores a class of phenomena among the most problematic for science – emotions. We have shown that the synthetic approach, when applied rigorously, produces an inter-individual view of emotional processes which heralds a generative two-way transmission of knowledge between epigenetic robotics and the psychology of emotional development.

## ACKNOWLEDGMENTS

This research is supported by the European Commission as part of the FEELIX GROWING project (<http://www.feelix-growing.org>) under contract FP6 IST-045169. The views expressed in this paper are those of the authors, and not necessarily those of the consortium.

## REFERENCES

- [1] D. Baird. *Thing knowledge*. University of California Press (2004).
- [2] L. Bich and L. Damiano. Theoretical and artificial construction of the living. *Origins of Life and Evolution of Biospheres*, 34:459-464 (2007).
- [3] G. Bocchi and M. Ceruti (Eds.). *La sfida della complessità*. Mondadori (2007).
- [4] M. A. Boden. *Mind as machine*. Oxford University Press (2006).
- [5] J. Bowlby. *A secure base*. Basic Books (1988).
- [6] V. Braintenberg. *Vehicles*. MIT (1984).
- [7] L. Cañamero. Emotion understanding from the perspective of autonomous robots research. *Neural networks*, 18:445-455 (2005).
- [8] L. Cañamero. Playing the emotion game with Felix. What can a Lego robot tell us about emotion? In: K. Dautenhahn, A.H. Bond, L. Cañamero, B. Edmonds. *Socially Intelligent Agents*. Kluwer (2002).
- [9] L. Cañamero and R. Aylett (Eds.). *Expressive characters for social interaction*. John Benjamins (2008).
- [10] L. Cañamero, A. Blanchard and J. Nadel. Attachment bonds for human-like robots. *International Journal of Humanoid Robotics*, 3:1-20 (2003).
- [11] J. Cassidy and P. R. Shaver. (Eds.). *Handbook of attachment*. The Guildford Press (1999).
- [12] A. Clark. An embodied cognitive science? *Trends in Cognitive Science*, 3, 9:345-351 (1999).
- [13] A. Clark and D. J. Chalmers. The Extended Mind. *Analysis*, 58:10-23 (1988).
- [14] L. Damiano. *Unità in dialogo*. Mondadori (2009).
- [15] L. Damiano and P. Dumouchel. Epigenetic embodiment. In L. Cañamero, P. Y. Oudeyer and C. Balkenius, *Epigenetic Robotics*, Lund University Cognitive Studies, 146: 41-48 (2009).
- [16] M. R. W. Dawson. From embodied cognitive science to synthetic psychology. In: *Proceedings of the First IEEE International*. (2002).
- [17] M. R. W. Dawson. *Minds and machines*. Blackwell (2004).
- [18] J. Doherty and J. Proust (Eds.). *Simulation and knowledge of action*. John Benjamins (2002).
- [19] P. Dumouchel. *Émotions*. Institut Synthélabo (1995).
- [20] P. Dumouchel. Sul modo d'esistenza degli strumenti scientifici: la conoscenza oggetto. *Il Protagora*, V, 12:327-332 (2008).
- [21] H. Foerster, von, and G. W. Zopf. (Eds.). *Principles of Self-Organisation*. Pergamon (1962).
- [22] V. Gallese. Intentional attunement. *Interdisciplines*. <http://www.interdisciplines.org/mirror/papers/1>

<sup>23</sup> See Footnote 15. See [10, 26-29]

- [23] E. Glasersfeld, von. *Radical constructivism*, The Falmer Press, London (1995).
- [24] E. Glasersfeld, von. The radical constructivist view of science. In A Riegler (Ed.). The impact of radical constructivism on science. Special issue of *Foundations of Science* 6(1–3). (2001).
- [25] E. Jantsch. *The Self-organizing Universe*. Pergamon (1980).
- [26] A. Hiolle, K. Bard and L. Cañamero. Assessing human responses to different robot attachment profiles. *Proceedings of the 18th International Symposium RHIC*, pp. 251-257 (2006).
- [27] A. Hiolle and L. Cañamero. Why should you care? An arousal-based model of exploratory behavior for autonomous robots. In *Proceedings of the 11th International Conference on Artificial Life*, 242-248, MIT (2008).
- [28] A. Hiolle and L. & Cañamero. Developing sensorimotor associations through attachment bonds. In *Proceedings of the 7th International Workshop on Epigenetic Robotics*, 45–52 (2007).
- [29] A. Hiolle, L. Cañamero and A. Blanchard. Learning to interact with the caretaker: a developmental approach. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interactions*, 425–436 (2007).
- [30] H. Keller. Attachment-Past and present. But what about the future? *Integr Psych Behav* 42: 406-415 (2008).
- [31] M. D. Lewis and I. Granic (Eds.). *Emotion, development and self-organization*. Cambridge University Press (2002).
- [32] P.L. Luisi. *The emergence of life*. Cambridge University Press (2006).
- [33] P. L. Luisi. The synthetic approach in biology. Manuscript. (2010).
- [34] H. Maturana. *Tutto ciò che è detto è detto da un osservatore*. In Thompson W. I. (Ed.). *Ecologia e autonomia*, 79-93, Feltrinelli (1988).
- [35] H. Maturana and F. Varela. *The Three of Knowledge*. Shimbhala (1987).
- [36] J. Nadel and D. Muir (Eds.). *Emotional development*. Oxford University Press (2005).
- [37] C. L. Neahianiv and K. Dautenhahn (Eds.). *Imitation and social learning in robots, humans and animals*. Cambridge University Press (2007).
- [38] R. Nunez R. and W. J. Freeman. (Eds.). *Reclaiming cognition*. Imprint (1999)
- [39] G. Pask. The Natural History of Networks. In M. C. Yovits and S. Cameron (Eds.), *Self-organizing systems*. 232-263, Pergamon (1960).
- [40] J. Piaget, *Biologie et connaissance*. Gallimard (1967).
- [41] R. Pfeifer, M. Lungarella and O. Sporns. The synthetic approach to embodied cognition. In P. Calvo and T. Gomila (Eds.). *Handbook of cognitive science. An embodied approach*. Elsevier (2008).
- [42] R. Pfeifer and C. Scheier. *Understanding intelligence*. MIT (2000).
- [43] I. Prigogine and I. Stengers. *La nouvelle alliance*. Gallimard (1986)
- [44] H. Radder (Ed.). *The philosophy of scientific experimentation*. University of Pittsburgh Press (2003).
- [45] A. Riegler. (Ed.). The impact of radical constructivism on science. Special issue of *Foundations of Science* 6(1–3). (2001).
- [46] H. A. Simon. *The sciences of the artificial*. MIT (1982).
- [47] I. Stengers. *Cosmopolitiques*, vol. 6, La Découverte (2003).
- [48] E. Tronick. *The neurobehavioural and socio-emotional development of infants and children*. Northon & Company (2007)
- [49] F. Varela. *Principles of Biological Autonomy*. North-Holland (1979).
- [50] F. Varela, E. Thompson and E. Rosch. *The embodied mind*. MIT (1991).
- [51] R. A. Wilson and A. Clark. How to situate cognition. In P. Robbins and M. Aydede (Eds.). *The Cambridge handbook of situated cognition*. Cambridge University Press (2009)

# Advancing Knowledge of Cognitive Development Sequences in Infancy

Frank Guerin<sup>1</sup>

**Abstract.** Human adults have relatively sophisticated cognitive abilities, and manage vast quantities of diverse knowledge. The amount of data in the genome, and the differences with other species, show that a relatively small amount of information must code a system that can bootstrap itself to this high level of sophistication. How this bootstrapping process works remains largely a mystery. The technique of computational modelling opens up the possibility of developing a complete model of this bootstrapping process which could allow us to understand the whole developmental sequence, starting with infancy. Existing computational models of infant development typically only model one episode in one area of competence, and these individual episodes have not been linked up; yet, the process by which new developments build on earlier achievements is central to the bootstrapping, and remains a mystery. This paper looks at what work needs to be done to take forward the idea of attempting to explain cognitive development at a level which could account for long sequences of development. We argue that in order to understand the developmental processes underlying longer sequences we first need to determine what these possible sequences are, in detail; existing knowledge of such sequences is quite sketchy. The paper identifies the need to discover a directed graph of behaviours describing all the ancestors of sophisticated behaviours. The paper outlines some experiments that may help to discover such a graph.

## 1 Introduction

Human cognitive development from infancy through to adolescence shows a fascinating dramatic increase in the sophistication of behaviours displayed. Arguments purely from the amount of data in the genome, and the differences with other species, show that a relatively small amount of information must code a system that can bootstrap itself to a high level of sophistication. How this bootstrapping process works remains largely a mystery; it must involve some process of incrementally building new knowledge on top of old knowledge structures. This “incremental building” is a process which is currently very poorly understood, and is therefore our focus of investigation. Computational modelling would probably be the most precise way to describe this development (if it is possible to reach such a level of precision). A computational model could make clear what learning mechanisms are in operation in cognitive development, and what mechanisms can account for multiple developments, building on each other.

In this paper we look at how we might advance our knowledge of the incremental building which happens during cognitive development in infancy, through studies involving both computational modelling and experimental psychology. The process starts with obser-

vations and theories from psychology. Computational modelling is necessary to test these theories and show up impossibilities and gaps in knowledge, where more precision is required to specify the developmental mechanism. This leads to the specification of further psychological experiments, and the cycle continues. The ultimate goal is to have a precise description of the mechanism which could account for some of the major cognitive developments which take place during infancy. The precision required is at the level of a running computer program, such that the program could be placed in a simulated world<sup>2</sup>, and could interact with the world, and develop cognitively, recapitulating the development of the infant. In this paper, as a first step in this process, we examine existing psychological results to see how useful they are for computational modelling, and what additional information would be desirable, and finally the type of psychology experiments which might reveal this information.

Computational modelling of cognitive development is nothing new of course; however there is no work which has attempted to model substantial sequences of development in detail. For example Schlesinger et al. [12] or Cohen et al. [5] model one part of development in some detail, while [6] models a longer sequences, but in very poor detail, i.e. Drescher’s simulated world was very simplified and precludes the modelling of many behaviours. One of the essential features of infant cognitive development is that it is *ongoing*; i.e., what is learnt in one learning episode is built on and forms the starting point for a subsequent learning episode. A computational account of the development must explain how one acquisition feeds into the next, in an ongoing sequence. This is a tough requirement to meet, and implies a large research programme, because each individual development modelled must be done with a view to subsequent developments, and must acquire knowledge structures in some representation which can feed into the next episode of learning.

In order to limit the scope of our research to a manageable chunk, the research programme we consider here is to be able to model the development from basic sensorimotor schemas such as Piaget’s sec-

---

<sup>2</sup> Rodney Brooks [2] has argued for the need to develop Artificial Intelligence (AI) with robots in the real world, rather than simulation. Brooks makes strong arguments for why relying on human introspection about how the world should be represented can lead to serious problems; this means that in building simulations we must be careful not to impose our representation on the AI system, but to expose it to data as might be seen by a vision system (for example), so that it solves the problem from the raw data, and does not have a human specified shortcut to higher representations. Rich Sutton’s verification principle [15] strengthens this point by pointing out that any human imposed shortcut could not be revisited by the AI system to be reformulated if necessary, hence the system would be brittle and restricted to scenarios the human had foreseen. However no strong argument has been made against simulation itself, provided it is done without “cheating”. There is no reason to believe that AI research (including computational modelling) cannot progress through work in simulations (past failures do not prove any general law).

---

<sup>1</sup> University of Aberdeen, Scotland, email: f dot guerin at abdn dot ac dot uk

ondary circular reactions (Piaget’s stage 3, beginning at roughly 4 months), right through to tertiary circular reactions (Piaget’s stage 5, running to roughly 18 months). The kinds of behaviours exhibited towards the start of this sequence would include striking objects, grabbing them, shaking them, etc. By the end we would see advanced behaviours such as retrieving a distant object with a stick. The aim is to create a computer simulation which could begin with the abilities at the start of this sequence, and could “play” in a simulated world, and with no further intervention from a programmer, could autonomously develop, through experiences in the world, to arrive at the behaviours marking the end of the sequence. Cognitive development in infancy involves both learning based on experience in the world, as well as the arrival of new competences through maturation of the nervous system. The focus of this paper is especially on the learning mechanism which builds knowledge based on experiences. However, our simulation will also need to be some modelling of maturation, which could take the form of a scripted sequence of developments which unfold at specific times.

The research agenda proposed in this paper would fall under the heading “developmental psychology”. The main input from Artificial Intelligence (AI) to this paper (and hence the link to this symposium’s theme) is *the very idea* of attempting to understand a developmental mechanism underlying long sequences of development. This is a very computational idea; i.e. that there could be a program, with a limited amount of initial information, which by gathering information from the environment, and modifying itself, can become very sophisticated. This computational perspective leads one to ask questions about long sequences of development, and the intermediate levels of competence, and training sequences that cause transitions. The questions thrown up by this attempt are questions which psychologists are not currently investigating.

Much of the developmental psychology literature about infancy probes the infant’s competences at various ages. The focus of the proposed investigation is on *development*, rather than competence at any particular age. (Assessing competence at particular ages is only interesting for the information it gives about the development which must have taken place between those ages.) We argue that there is a need to conduct research on infancy which identifies ordered sequences of intermediate competences, and the training experiences which lead to development from one intermediate level of competence to the next. Research to identify training experiences which can lead to development has been carried out with older children, but very rarely with infants.

## 2 Requirements: Precise Models of Infant Cognitive Development

The ultimate aim of the research which this paper is concerned with is to have a complete mechanistic description of infant cognitive development; i.e., working computer model, which could recapitulate the development of a human infant. Various levels of detail can be envisaged; the most detailed being a realistic infant-like robot which interacts with real human caregivers, while simpler versions could include simulated robot infants in simplified simulated worlds which could recapitulate some of the main themes of development. This paper considers the simpler end of this spectrum, and focuses on just some of the developments which take place. We are particularly interested in developments relating to fine motor skills rather than gross motor skills. This section of the paper looks at the shortcomings of existing work in psychology, and looks at what knowledge may be needed from further psychological investigations in order to guide

the development of a computational model.

### 2.1 Shortcomings of Piaget’s Theory

Piaget’s theory would seem to be useful for informing the construction of computational models of infant development, because it describes a learning function which operates at all ages, and which constructs new knowledge structures. There are some problems with it however, as described in this section. Piaget’s theory groups the behaviours of the first two years into six sensorimotor stages [10]. In explaining his theory about the development between the different stages he tends to describe the development in general terms, rather than describing how individual behaviours belonging to a particular stage lead to some other behaviours from the next stage. There are a number of problems with this.

Firstly it is not certain that it is sensible to group the behaviours into the six stages which he groups them in. Take for example Piaget’s fifth sensorimotor stage. In this stage he group together (1) the behaviour of the support (e.g. pulling a towel in order to bring an object resting on it closer); (2) the behaviour of the string (pulling a string which is attached to an object, in order to bring this object closer); (3) the behaviour of the stick (using a stick to retrieve a distant object). These behaviours tend to occur at quite different ages, the behaviour of the stick in particular is much more difficult than the string or the support, and can occur four months later. Furthermore, many of these behaviours have a number of different levels of competence within them, for example the support can occur in the fourth sensorimotor stage as a means-end behaviour, without full understanding of the situations in which it will work; the string can also occur in a similar fashion, and there can be a considerable gap between competence on the string on a horizontal surface and vertically [16]. A second example is the objective knowledge of space, which Piaget groups with the fifth sensorimotor stage. This clearly will not arrive all together, but many fragments of objective knowledge must be built up during this stage, and this detail is not described. This suggests that Piaget’s grouping is at best too coarse grained (and therefore not very useful), or at worst mistaken and misleading. The questionable utility of sensorimotor stage 5 is further illustrated by considering the results of an acceleration study by Wishart and Bower [18]. In this study the viewing of training displays accelerated infants’ development of the stage 5 behaviour of correctly retrieving hidden objects; however it is extremely doubtful (although the tests were not carried out) that these infants would be capable of other stage 5 behaviours, such as experimental discovery of new means (one of the hallmarks of stage 5). This calls into question the utility of Piaget’s grouping because some stage 5 behaviours can develop seemingly independently of others. (Piaget’s stage 5 has just been used as an example here, similar comments could be made for stages 3 and 4, which also pack in quite a variety of different behaviours.)

Secondly, the frequency of behaviours exhibited from each Piagetian stage take the form of overlapping waves, much like Siegler’s theory of development [3, see p. 7]. This is roughly sketched in Figure 1; one can see that at a particular age (say one year) an infant will display behaviours from a number of different stages. As a child enters a new stage the behaviours of that stage gradually start to become more frequent, but those of the previous stages may not decline appreciably for quite some time. This means that Piaget’s theory does not make it clear which behaviours necessarily precede which others. For example a late stage 4 behaviour may well first occur after an early stage 5 behaviour, furthermore Piaget does not go into details

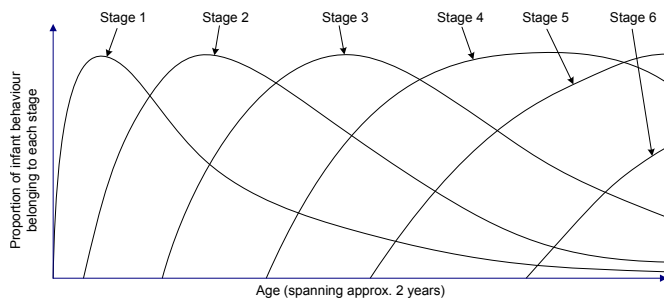


**Figure 1.** Piaget's sensorimotor stages as overlapping waves.

about the necessary order of most behaviours within a stage. This is problematic because we would ideally like to know the ordered sequence of development for each individual behaviour and intermediate level of competence. Uzgiris and Hunt [16] proposed an alternative categorisation which suggested instead six parallel tracks of development, such that development on one track could proceed faster (or slower) than on other tracks. The order of acquisitions within any track cannot change, but between tracks, variations in relative order are possible, and this is where Uzgiris and Hunt can differ from Piaget; Piaget lumps a number of acquisitions together in one stage, and does not view them as independent tracks. There is evidence to support Uzgiris and Hunt's view [16, p. 134 and p. 138].

Given the question marks over the suitability of Piaget's grouping of the behaviours in his sensorimotor stages, it is problematic that he tends to mostly describe development in general terms. For example, in his description of how stage 5 develops from the infant's earlier knowledge, he describes how the new ability to experiment is derived from the ability to recognise the results of experimental acts, because these results match schemas learnt previously [10, Chap. 5, Sec. 1, especially p. 276-277]. However, he does not describe which earlier schemas in particular permit the recognition of which results, thus facilitating a particular type of experimentation. Thus Piaget has a sort of grand unified general theory of development, without having a detailed account of how some specific behaviours develop. This is a little suspect, as one would like to see a grand unified theory built on a foundation of detailed knowledge of specific instances. Piaget's general theory may or may not be accurate, but certainly we will need more detail before we can make this judgement, and Piaget's evidence falls far short of what would be required to support his theory as a credible explanation of all of infant development.

For the work proposed in this paper we would like to postpone the formulation of a general theory, and instead focus exclusively on trying to explain in detail the mechanism which can account for some specific acquisitions. When a great deal of this specific research is done it may be possible to see the big picture, and compare with Piaget's theory. It is likely that Piaget's description will eventually turn out to be accurate for certain developments, but that much more is needed to explain other aspects. Piaget's genius may lie in the fact that he was able to see the outline of a general theory without needing to know all the details.

Note that the shortcomings of Piaget's theory described here were found by taking a computational perspective on the problem. The computational perspective leads one to think about how to formulate a learning mechanism in precise terms, which in turn leads one to realise how hopelessly vague Piaget's description of stage transition

can be. This leads to the effort to find concrete precise descriptions of transitory behaviours, in order to give sufficient data to guide the search for a computational mechanism which could account for the transition. Thus it is an example of how adopting an AI perspective can lead to new research programmes which may answer some of the core issues which concern developmental psychologists (in this case the search for a mechanistic explanation for cognitive development).

## 2.2 Shortcomings of More Recent Psychology Research

With regard to more recent work in infant cognitive development, there is a lot of work on probing infants' competences at various ages (for example [1]), but very little work addresses the issue of *development*, and explains what leads to the changes in competences. Chen and Siegler [3] have noted this gap, and also note a contrast between the types of research carried out on infants' and toddlers' thinking when compared with research on the thinking of older children. With older children, we also see research on competences, but in addition, we see "studies aimed at revealing mechanisms of change" [3]. Part of the reason for this difference is the different experimental paradigms used. For example, with older children it is common to ask them for explanations for their decisions; obviously these techniques do not translate easily to infancy research. However, Chen and Siegler are of the opinion that the gap between the two types of research is unnecessary, and many techniques applied to older children can work with infants (some of these techniques are described in Section 3 below).

Chen and Siegler advocate closing the gap with new experiments on infants, using appropriate experimental paradigms, to answer questions such as "Through what process of change do children become able to do X". To this end they have conducted a study on children as young as 18 months to analyse the strategies they use to retrieve a distant object, when provided with an array of tools (some useful and some not). The studies we advocate however will need to experiment with younger infants, as we want to examine how skills such as tool use are first acquired. The methods of Chen and Siegler will be very useful for our proposal.

## 2.3 Requirements for Machine Learning

From a computing perspective, the task of creating a computational model of infant cognitive development could be seen primarily as a machine learning problem. To define the machine learning problem we need to specify the tasks to be performed, and the training experience. The tasks are simply the types of behaviours which an infant is capable of at 18 months, and incapable of at 4 months say. For example, one can use some of the types of tasks documented by Piaget, which 18 month-olds are typically capable of, such as the behaviour of the stick (retrieving an out-of-reach object by means of a long stick). The training experience is varied and complex, and most of the training period is spent attempting other intermediate tasks, some of which lead to learning which is useful for performing the 18 month tasks. For this reason it may make more sense to look at a sequence of separate machine learning problems, each with its own training experience, and target tasks, and where the learning on earlier problems in the sequence facilitates later learning. What the computer modeller needs from psychology then is information about a sequence of these learning problems. Each of these learning problems will involve the transition from some competences at a particular age to a more advanced version at a later age, and ideally the

gap between these two ages should not be too large, so that we are modelling small steps in development. Psychology experiments will need to find the sequence of competences, and the training experiences which lead to developments. For computer modelling we also need to decide on what sensorimotor abstractions and representations to use. This is much more difficult to determine from psychological experiments, but it may be sufficient to simply guess, and refine those guesses as needed in order to be able to model a whole sequence of developments. The more detailed our knowledge of the sequence of behaviours and the training experiences is, the more well informed our guesses about representations will be.

## 2.4 The Graph of Development

An analogy between evolution and cognitive development helps to explain the idea here. In looking at the many organs which different animals have evolved it is often difficult to imagine the path along which they evolved. The organs are quite complex (for example the eye), and if one were to guess at the mechanism of evolution based on scant evidence, one might conjecture a much more sophisticated mechanism than necessary. However, if one sees all the precursors in the fossil record for an organ, such as the eye, then the development pathway becomes more clear, and the mechanism which could account for the evolution can be simpler. Similarly, in looking at the sophisticated abilities of a toddler, and the large gap between those and a young infant's abilities, one might conjecture an overly complex learning mechanism to account for it; however, if the path of development can be traced out in detail, showing small developments between intermediary abilities, then the development might be explained with a simpler mechanism.

In evolution there is a tree, and each species (or organ) can be an ancestor for one or more others. In cognitive development the analogy of the species (or organ) is a knowledge structure, which could be a schema in Piagetian terminology, or could be a concept or skill or sensorimotor abstraction. Each knowledge structure may have multiple immediate ancestors (unlike evolution) because new structures are often built from relationships among existing ones (see for example Cohen's theory [5]). This means that in place of evolution's tree, cognitive development has a *directed acyclic graph*. The nodes of the graph represent knowledge structures, which could include for example a sensorimotor abstraction, allowing recognition of some situations where an action has a certain effect, or a motor control policy allowing performance of a skilled action. The directed edges of the graph have the meaning that the set of immediate ancestors of a node are the necessary and sufficient precursors of that node. The graph can then be supplemented with a new type of node to describe the training experiences that can lead to developments, i.e. a new training experience node can be inserted between a set of ancestors, and the descendent which they can give rise to. Finally we would specify machine learning algorithms which can start with the knowledge of some ancestor nodes, and use the training experience to generate the descendent nodes.

The above description of a development graph is a sort of gold standard, which could serve as a long-term goal for research in the field. In the medium term it is probably unreasonable to pursue it given our present poor state of knowledge about infant cognition. In particular we know very little about knowledge structures existing at various stages of development, and all we can directly observe are the behaviours which the infant is capable of. For this reason it may be more feasible to attempt to find a graph of behavioural milestones. This would again be a directed acyclic graph, but each node

describes observable behaviour; i.e., a certain behavioural milestone, or training experience (these no longer need to be different types of nodes as all nodes describe behaviour now). The directed edges of the graph have the meaning "is a necessary precursor"; i.e. the later behaviour should not be possible if the infant is incapable of the earlier behaviour. We would not claim that the immediate ancestors of a node are sufficient to enable the later behaviours, as this is a graph that will need to be built up incrementally; we can test behaviours to determine if they are necessary as ancestors, but we will not know if we are missing some.

Some additional factors complicate the infant development graph. Firstly, not all infants develop along the same paths, thus there are some later behavioural milestones (say milestone  $m$ ) which have more than one set of possible immediate ancestors. This could be handled by having the behavioural milestone  $m$  represented by more than one node in the graph, i.e., once for each unique set of possible ancestors which could lead to it. The various nodes representing the milestone could then link to a single one, so that this would be a unique ancestor on which later behaviours could build. In any event, the issue of alternative paths need not be a great concern for initial investigations, because it may suffice to discover one possible developmental progression. The second complicating factor is that not all new behaviours are derived from others; some arise due to maturation, or due to a combination of maturation, and a certain level of development having been attained. This means we may include nodes in the graph which do not have ancestors or which are not arrived at via training experience.

For each intermediate competence, and each node describing a training experience, we need a fairly precise description of the behaviour, in order to avoid ambiguity. We must describe standard objects to be used, and detail the movement expected of the infant. For example, in a retrieval behaviour using a stick we must specify the dimensions of the stick (as the behaviour with a short stick is much easier than a long one for a more distant object), we must also specify the rigidity and weight of the stick; for the object to be retrieved we must specify the shape, dimensions, weight, colour, and frictional coefficient with the supporting surface (an object that rolls or slides easily is more difficult to retrieve, as the infant's poorly controlled swipes will put it out of reach); for the surface we must also specify the pattern (as it may be that a pattern with some visual landmarks is used to help interpret relative motion); for the remainder of the environment we should specify any other objects in the scene, and what background is visible in the room.

Note that this call for precise specification is not intended to create an artificial micro-world in which problems can be solved without tackling the full complexity that is entailed by the variation in the real world. The precise specification is necessary so that the same experiments can be repeated with different infants, and so that ambiguity will not result in another experimenter using different materials and unwittingly testing a qualitatively different behaviour. We are here describing experiments for infants, who naturally have a strong ability to generalise, and handle minor variations in materials and conditions. If such a detailed specification of behaviours could be made for the graph, then it would also be extremely useful for experiments in developmental robotics. Again, this is not an attempt to define an artificial micro-world, but rather to show a sequence of behavioural competences and training experiences which lead from one to the other, which form a potential developmental trajectory, which roboticists could attempt to follow.

In summary, we may not yet be ready to formulate a complete theory of infant development, explaining the knowledge structures

built, and the mechanism which builds them; however, as a first step, it may be useful to come up with a directed graph which shows which behaviours build on which others. Such a graph would be useful for computational modelling efforts, as it sets the sequences of tasks which need to be modelled, along with the training experience, and it gives insight into the incremental learning required; i.e. how early learning transfers to later learning. Computational modelling would then begin the process of guessing at the types of knowledge structures which are responsible for the competences shown at each node of the graph, and the learning algorithms which can build these knowledge structures, given the appropriate training experiences.

#### 2.4.1 Potential Criticism of the Approach

The development graph we propose may feel like a bit of a step backwards for two reasons. Firstly it is a little behaviourist, in that we are relying solely on observable behaviour as guidance for computational modelling, even though a number of works in the psychological literature have already made conjectures about knowledge possessed by infants at various ages (see for example [1]). There is however considerable controversy about the claims made about infant knowledge [4, 7, 8], so it may be safer to rely on observable evidence from infants acting, and then to let the computer modelling come up with the simplest models which could account for the behaviour observed. Much of the work which purports to show advanced competences in infants might be criticised in the same way that a lot of early artificial intelligence work was criticised by Brooks [2]. Researchers may be too quick to assume an adultlike mind within the infant. This is simply a case of transferring the adult's model of his/her own thinking (known only by introspection) across to the infant; but this is by no means the simplest explanation for an infant competence, and science should always prefer the simplest explanation. Ascribing an adult world model to the infant may seem like the simplest explanation if one overlooks the enormous complexity of adult reasoning. A further point is that, as adults, we might well not be solving problems in the way we think we are solving them. We may be in fact solving the task subconsciously in the same way as we learnt during infancy, and then fitting an adult rationalisation on top of the solution after the fact. This means that it may be really a long time before an infant has an adultlike concept of object permanence for example, or perhaps never (if the adult's idea of the concept from introspection is not in fact what the adult is reasoning with). In summary we should not presume what is inside the infant's head based on adult introspection, we should just make the simplest model that can account for the behaviour, and adjust it whenever it is inadequate.

The second reason why our proposal may seem a bit of a step backwards is that we are proposing to make no assumptions about the general features of the mechanism of development, even though theories such as Piaget's have made conjectures about this mechanism. For example, if following Piaget's theory we would perhaps explicitly focus on modelling his sensorimotor stages, and the transitions between them. However, given the concerns described about his theory above, we believe that it may be better to start computational modelling from scratch, guided purely by evidence from behaviour. This process begins with some of the infant studies already carried out, such as Willatts's [17], discussed in the next section.

#### 2.4.2 Work Already Done

This section briefly looks at some existing work which gives some fragments of a development graph. The work of Piaget [10, 11] has identified a partially ordered sequence of behaviours which build on each other. Much of Piaget's writing on this is speculative, and experiments have yet to be carried out to test many of his specific conjectures. An example of one of his conjectures is as follows. Piaget's son Laurent acquired the ability to discard one object in order to grasp another, at about 7 months, 29 days [10, Obs. 125]. Piaget conjectured that this behaviour derived from the ability to set aside an obstacle which prevented him from grabbing a desired object (this behaviour having been acquired just previously). This idea of one behaviour being derived from the other does not seem to have been investigated subsequently (and there are many more such examples yet to be investigated).

The work of Uzgiris and Hunt [16] (mentioned above) provides more fine grained detail, and the beginnings of a development graph, with six parallel branches. This defines a partial ordering of behaviours, and is what we require, except that we need to know more about behaviours intermediate between those they describe, and the training experiences which go between them. (Uzgiris and Hunt's interest was more in devising tests, rather than gaining insight into the mechanism of cognitive development.)

For determining training experience, Willatts's work on means-end behaviours [17] provides excellent evidence for the training experience which can lead an infant to knowledge of the means-end action which can retrieve an object supported by a towel. The above works are all useful steps on the way towards the directed graph which we envisage. These works however represent exceptions, as the main thrust of research in cognitive development is more concerned with determining competence's at specific ages. We believe that those works which specifically identify developments which lead to other later developments give some of the most useful knowledge to inform computational modelling efforts.

### 3 Promising Experimental Paradigms

This section briefly looks at the experimental paradigms which are most promising for taking forward the research agenda we are concerned with.

#### 3.1 Testing of Sensorimotor Skills

Simple tests of behaviour will be an essential part of any experiments to discover the graph of development. Tests can simply involve observing the spontaneous behaviour of the infant with some objects, or setting the infant a task, such as to retrieve a toy which interests the infant. There are a great number of examples of these tests in Uzgiris and Hunt's book [16]. Once an infant can retrieve a toy in a situation, further variations can be used to probe the infant's competence in more detail. For example, if the infant can retrieve an out of reach toy placed on a towel, by pulling the towel, then the experimenter can hold the toy just above the towel and see if the infant continues attempting to pull the towel. Some experiments are also done by first demonstrating a procedure to the infant, and then examining if the infant is capable of imitating the procedure.

Willatts [17] introduced a further technique in tests by not only recording success or failure on the task, but also monitoring the infant's gaze, in order to have an objective measure which could discriminate between accidental success, or intentional success. This

was used on the task of pulling a towel to retrieve a supported toy. Younger infants (about six months) tend to give up on the toy and play with the towel instead, but in doing so they often accidentally bring the toy into reach. Willatts was able to monitor the infant's gaze, and to show that there was a transition: whereas the younger infants gave up on looking at the toy, as they got older, there were more glances towards the toy, suggesting that pulls of the towel were intentional in order to retrieve the toy.

### 3.2 Intervention to Accelerate Development in Longitudinal Studies

An intervention which can accelerate development of one group of infants relative to a control is one of the clearest ways to identify the experience which can lead to a development (which is required in our development graph). In a longitudinal study by Wishart and Bower [18] a group of infants was trained by watching a series of displays at regular sessions over a number of weeks, and tracking the objects seen in the displays. There were three displays: (1) an object which moved over a platform, i.e. coming into contact with the platform, and moving on past it; (2) an object that moved behind a screen and re-emerged from the other side; (3) an object which moved through a hollow cylinder and re-emerged. These infants were then tested on standard Piagetian object retrieval tasks, using the classic "A not B" problem. The tasks included an object placed on platforms, an object hidden behind screens, and an object hidden under a cup. These tasks belong to Piaget's fifth sensorimotor stage. The infants showed a dramatic acceleration in development, and succeeded on all three tasks well before the average ages at which success is typically achieved on these tasks. This study then very clearly shows the type of training experience which can lead to development from one level of competence to another in a very specific task. Such studies are resource intensive as infants must be studied over a relatively long period of time, however they give probably some of the best evidence which could be used to find the development graph of Section 2.4. The basic idea of accelerating development has also been used over shorter timescales in the microgenetic studies of Siegler and colleagues as described next.

### 3.3 Microgenetic Studies

Siegler has pioneered the microgenetic method [14]. The method is defined by three characteristics (reproduced from [3]):

1. an observation period spanning the time from the beginning of the period of rapid change to the stable use of target ways of thinking;
2. a high density of observations during this period, relative to the rate of change; and
3. intensive, trial-by-trial assessments of ongoing changes, both qualitative and quantitative.

The method seems to have rarely been used on infants, apart from the study of Chen and Siegler [3] which went as low as 18 months of age. By looking at some studies on older children, we can get an idea of how we might translate the same ideas to infancy. Siegler and Chen [13] have carried out a study to probe children's reasoning about how much water will be displaced by various objects. This study ran over two consecutive days, children were tested at the start and the end, and underwent training in the middle. Part of the training was designed to accelerate development; some of the children were asked to explain only their own answer, whereas others were asked

to explain correct and incorrect answers; this latter group were more likely to notice the roles of the variables in the problem, and more of them developed to the stage of correct answers within the period of the study. The general pattern of development here could translate to infancy experiments: noticing new variables, formulating more advanced approaches that incorporate the new variables, and generalising discoveries to new problems. For example in an infant tool use experiment, the infant must notice the position of the end of the tool relative to the infant's hand; younger infants will overlook this, but older infants know how to orient the tool, for example to maximise reach. A microgenetic study could experiment with various training experiences to determine which will lead the infant to pay attention to this variable. We cannot ask the infant for explanations, but we can expose the infant to situations where distracting aspects are removed, and it is more clear that the relative position of the end of the tool is the important variable across trials.

In looking at Chen and Siegler's study on tool use [3], one can also see methods that could be employed in studies of younger children. Chen and Siegler have outlined five steps in the acquisition of new strategies (reproduced from [3]): (1) acquiring the strategy of interest, (2) mapping the strategy onto novel problems, (3) strengthening the strategy so that it is used consistently within given types of problems where it has begun to be used, (4) refining choices among alternative strategies or alternative forms of a single strategy, (5) executing the strategy of interest increasingly effectively. In their study they looked at strategies for retrieving a distant object placed on a table when tools were available. Strategies used by the children included reaching across the table, or climbing on it, using a long tool, or turning to the parent for help. The same methods might be used where strategies are more fine grained, as required for our purposes. For example one could analyse the strategies that can be used with a long tool, during the phase where the infant is not yet competent in its use. In this case strategies will include: directing the stick towards the object; hitting the object; getting the end of the stick behind the object; repeated hitting on one side to bring gradually closer; changing the side of hitting when the object moves too far to one side.

Combining the experimental paradigms described here, a complete research methodology for determining the development graph could proceed as follows:

1. Initial longitudinal studies to graph a sequence of developments, with microgenetic studies to observe transition periods.
2. Formulation of hypotheses about which behaviour acquisitions may be used by later acquisitions.
3. These hypotheses could then be tested via studies which attempt acceleration of developments.

## 4 Some Specific Sub-Graphs to Investigate

The three Piagetian stage 5 behaviours mentioned above (i.e., string, support, stick) could be a starting point for investigating sub-graphs of the complete development graph. Each of these behaviours is a node which forms the endpoint of a subgraph containing all behaviours which lead to it. For each of the behaviours we already could sketch an outline of a sequence of developments leading to it.

For the behaviour of the stick we can sketch an outline sequence of some behavioural milestones as follows (not all behaviours here are in a strict ordering).

- The ability to grab an object;

- The ability to manipulate the object, to turn it and grab different parts (especially when this is applied to long objects, such that the near part can be pulled in order to bring the far part into range for detailed investigation);
- The ability to repeat the action of hitting a grabbed object against a surface to produce a noise (if discovered by chance);
- The ability to hit a grabbed object against another stationary object to produce a sound (if discovered by chance);
- The ability to intentionally grab an object in order to use it to hit another object (at this stage the grabbed object is considered to be located only at the fist, even if a long object);
- The ability to hit the other object with the end of the (long) grabbed object, when the long object happens to have been grabbed in such a way that it already extends forward;
- The ability to direct the end of an object (i.e. an object with a rod-like shape) towards another (this entails The ability change the angle of an object, relative to the hand);
- The ability to use a short rod as an extension of the hand, to manipulate rattles, blocks, etc.;
- The ability to use a short rod to retrieve objects just out of reach, by a single motion which “scoops” the object into reach;
- The ability to release a rod and re-grasp it in a better place, in order to extend the reach;
- The ability to retrieve with a longer rod;
- The ability to use a long stick to bring a distant object by a series of successive hits to each side of the object. (This is quite advanced compared to the behaviours above, and many intermediates between them need to be found.)

Existing knowledge of these intermediate abilities is quite sparse, and many studies remain to be carried out to find relative orderings and the training experiences to cause transitions. The specification for each test of intermediate behaviour must be quite detailed, to specify the conditions under which success can be achieved or not. Even the final competence of our sequence will not have the same capability to retrieve objects in tricky situations, when compared with an adult using the same stick; therefore the behaviour test must make clear exactly the minimum requirement to pass.

The sequence of behaviours for the stick above is mostly a straight ordering, and is missing the branching out that we expect to occur as we trace through the ancestors. Some of Piaget’s general hypotheses about stage transition can be used to guide the search for further necessary behaviours in this sub-graph; we can test if the general hypotheses hold true in the specific behaviours we analyse. For example, in the fifth sensorimotor stage there is an ability to recognise little motions caused by chance (e.g. when the stick hits the object and it moves a little closer, or a little further away). Piaget hypothesises that these recognitions are possible because of schemas that were learnt during the third and fourth stages. In fact Piaget does give a little more detail on what he suspects is the developmental sequence in these specific behaviours. In the retrieval with a long stick, the recognition of a movement closer or further away is hypothesised to rely on understanding of these motions which were gathered during experience with displacing objects by the support and the string [10, p. 302, top]. The understanding of the motion during acquisition of the behaviour of the support is hypothesised to rely on the behaviour of moving objects by means of a string during the third sensorimotor stage [10, p. 288]. Obviously not all infants need to learn the support and the string before graduating to the stick; in the case where the stick is learnt without the other two, the interpretation of motion (produced fortuitously) must rely on knowledge of rela-

tionships learnt earlier, probably in the fourth sensorimotor stage, such as perhaps the relationship “in front”. The basic understanding of one object being in front of another, and removing an occluder is rather simple and acquired long before the behaviour of the (long) stick. There is likely a relatively long developmental sequence to be worked out here, which could be explored by suitable tests to determine which motions can be distinguished by the infant, and recognised as leading to an object coming closer, and into reach.

This section has just outlined some behaviours that could be investigated in order to produce a graph of development. The main point of the paper is that this detailed graphing is essential in order to discover the mechanism of cognitive development via accurate computational modelling, i.e., to find the learning function which builds new knowledge structures. If we do not have detailed knowledge of behavioural sequences, then we are very unlikely to chance upon the actual mechanism in use by infants, and a computational modeller is more likely to insert inappropriate knowledge and representations, and to be misled by introspection, and the way we expect infants should reason.

Discovering a mechanistic description of cognitive development (via computational modelling) is a contribution to research in biology, because it is giving a detailed understanding of a biological system. The methods to be used will come from AI, as AI has already done a lot of work in machine learning and various representations, which can be borrowed for the modelling task.

## 5 Infancy Studies as a Gateway to Complete Models of Development

Infancy research holds the promise of helping to advance the larger enterprise of understanding cognitive development in humans at all ages. Infancy research may help us to gain insight into the mechanism of knowledge building at an age where it may be easier to guess at what builds on what, due to the limited set of possibilities in infancy (relative to other ages). When modelling development on humans older than infants one can only model a very small portion of the knowledge of the human, and so tasks involving selecting appropriate knowledge are overly simplified.

Reasoning by analogy is one useful example to illustrate this. Analogy is recognised to be an integral part of human cognition, but studies which investigate it [9, for example] cannot model all the knowledge which an agent has. The typical task in analogy studies is to be given an example related pair of objects “ $x$  is to  $y$ ” and to be given a third object  $p$ , the subject must then find a fourth object  $q$  such that  $p$  is to  $q$  as  $x$  is to  $y$ . However, the object  $q$  is typically selected from a small handful of possible candidates, sometimes as low as two. This is clearly missing a major part of the problem facing a human with a more complete set of knowledge. The human must select from a vast array of possible candidates, and must choose a representation which is appropriate to the problem. At the present time one could hardly hope to model all the knowledge of an adult, and thus solve the complete problem including selection from many alternatives. However, it may be feasible to model all the fine object manipulation abilities of a young infant, and to build on these, so as to have a reasonably complete model of the infant’s knowledge; analogical tasks (such as finding analogous objects or parts of objects for manipulation) could then be studied in a more complete system.

## 6 Acknowledgments

Thanks to the reviewers' comments this paper was revised substantially to point out more clearly its connection with the theme of how AI can contribute to research on biological systems.

## REFERENCES

- [1] Renee Baillargeon, 'Infants' reasoning about hidden objects: Evidence for event-general and event-specific expectations', *Developmental Science*, **7**(4), 391–424, (2004).
- [2] R. A. Brooks, 'Intelligence without representation', *Artificial Intelligence*, **47**, 139–159, (1991).
- [3] Zhe Chen, Robert S. Siegler, and Marvin W. Daehler, 'Across the great divide: Bridging the gap between understanding of toddlers' and older children's thinking', *Monographs of the Society for Research in Child Development*, **65**(2), i–105, (2000).
- [4] L. B. Cohen and C. H. Cashon, 'Infant perception and cognition', in *Comprehensive handbook of psychology. Volume 6, Developmental Psychology. II. Infancy*, eds., R. Lerner, A. Easterbrooks, and J. Mistry, 65–89, New York: Wiley and Sons, (2003).
- [5] L. B. Cohen, H. H. Chaput, and C. H. Cashon, 'A constructivist model of infant cognitive development', *Cognitive Development*, **17**, 1323–1343, (2002).
- [6] G. L. Drescher, *Made-Up Minds, A Constructivist Approach to Artificial Intelligence*, MIT Press, 1991.
- [7] M. M. Haith, 'Who put the cog in infant cognition: Is rich interpretation too costly?', *Infant Behavior and Development*, **21**, 167–179, (1998).
- [8] Iain Jackson and Sylvain Sirois, 'Infant cognition: going full factorial with pupil dilation', *Developmental Science*, **12**(4), 670–679, (2009).
- [9] Robert Leech, Denis Mareschal, and Richard P. Cooper, 'Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill', *Behavioral and Brain Sciences*, **31**(4), 357–378, (2008).
- [10] J. Piaget, *The Origins of Intelligence in Children*, London: Routledge & Kegan Paul, 1936. (French version published in 1936, translation by Margaret Cook published 1952).
- [11] J. Piaget, *The Construction of Reality in the Child*, London: Routledge & Kegan Paul, 1937. (French version published in 1937, translation by Margaret Cook published 1955).
- [12] M. Schlesinger, D. Amso, and S.P. Johnson, 'Simulating infants' gaze patterns during the development of perceptual completion', in *Proceedings of the Seventh International Conference on Epigenetic Robotics*, Piscataway, NJ, USA, pp. 157–164, (2007).
- [13] R. S. Siegler and Z. Chen, 'Differentiation and integration: Guiding principles for analyzing cognitive change', *Developmental Science*, **11**, 433–448, (2008).
- [14] R. S. Siegler and K. Crowley, 'The microgenetic method: A direct means for studying cognitive development', *American Psychologist*, **46**, 606–620, (1991).
- [15] Richard S. Sutton. Verification, the key to ai, 2006. Unpublished document, available on author's webpage <http://www.cs.ualberta.ca/~sutton/IncIdeas/KeytoAI.html>.
- [16] Ina C. Uzgis and J. McV. Hunt, *Assessment in infancy : ordinal scales of psychological development*, Urbana : University of Illinois Press, 1975.
- [17] P. Willatts, 'pulling a support to retrieve a distant object', *Developmental Psychology*, **35**, 651–667, (1999).
- [18] J. G. Wishart and T. G. R. Bower, 'A longitudinal study of the development of the object concept', *British Journal of Developmental Psychology*, **3**, 243–258, (1985).

# Modeling visual Affordances: The Selective Attention for Action Model (SAAM)

Christoph Böhme\* and Dietmar Heinke<sup>1</sup>

## Abstract.

Classically, visual attention is assumed to be influenced by visual properties of objects, e. g. as assessed in visual search tasks. However, recent experimental evidence suggests that visual attention is also guided by action-related properties of objects (“affordances”)[1, 2], e. g. the handle of a cup *affords* grasping the cup; therefore attention is drawn towards the handle. In a first step towards modelling this interaction between attention and action, we implemented the Selective Attention for Action model (SAAM). The design of SAAM is based on the Selective Attention for Identification model (SAIM)[3]. For instance, we also followed a soft-constraint satisfaction approach in a connectionist framework. However, SAAM’s selection process is guided by locations within objects suitable for grasping them whereas SAIM selects objects based on their visual properties. In order to implement SAAM’s selection mechanism two sets of constraints were implemented. The first set of constraints took into account the anatomy of the hand, e. g. maximal possible distances between fingers. The second set of constraints (geometrical constraints) considered suitable contact points on objects by using simple edge detectors. We demonstrate here that SAAM can successfully mimic human behaviour by comparing simulated contact points with experimental data.

## 1 Introduction

Actions need to be tightly guided by vision in our daily interactions with our environment. To maintain such a direct guidance, J. J. Gibson postulated that the visual system automatically extract “affordances” of objects [2]. According to Gibson, affordance refers to parts or properties of visual objects that are directly linked to actions or motor performances. For instance, a handle of a cup *affords* directly a reaching and grasping action. Recently, experimental studies have produced empirical evidence in support for this theory. Neuroimaging studies showed that objects activate the pre-motor cortex even when no action has to be performed with the object [4, 5]. Behavioural studies indicated response interferences from affordances despite the fact that they were response-irrelevant [6, 7]. For instance, a recent study in Ref. [8] demonstrated that pictures of hand postures (precision or power grip) can influence subsequent categorisation of objects. In this study, participants had to categorise objects into either artefact or natural object. Additionally, and unknown to the participants, the objects could be manipulated with either a precision or a power grasp. The study showed

that categorisation was faster when the hand postures were congruent with the grasp compared to hand postures being incongruent with the grasp. Hence, the participants’ behaviour was influenced by action-related properties of objects irrelevant to the experimental task. This experiment together with earlier, similar studies can be interpreted as evidence for an automatic detection of affordances.

Interestingly, recent experimental evidence suggests that not only actions are triggered by affordances, but also that selective attention is guided towards action-relevant locations. Using event-related potentials (ERP) Handy *et al.* showed that spatial attention is more often directed towards the location of tools than non-tools [9]. Pellegrino *et al.* present similar evidence from two patients with visual extinction[10]. In general visual extinction is considered to be an attentional deficit in which patients, when confronted with several objects, fail to report objects on the left side of their body space. In contrast, when faced with only one object, patients can respond to the object irrespective of its location. This study demonstrated that this attentional deficit can be alleviated when the handle of a cup points to the left. Pellegrino *et al.* interpreted their results as evidence for automatically encoded affordance (without the patients’ awareness) drawing the patients’ attention into their “bad” visual field.

This paper aims to lay the foundations for a computational model of such affordance-based guidance of attention. We designed a connectionist model which determines contact points for a stable grasp of an object (see Fig. 1(a) for an illustration). The model extracts these contact points directly from the input image. Hence, such a model could be construed as an implementation of an automatic detection of object affordances for grasping. To realise the attentional guidance through affordances, we integrated the selection mechanisms employed in the Selective Attention for Identification Model (SAIM)[3]. Since this new model performs selection for action rather than identification, we termed the new model Selective Attention for Action Model (SAAM). There are only few computational models of affordance[11, 12]. However, Faggs *et al.* model does not process multiple-object scenes. On the other hand Ciseks model considers attentional processing with respect to pointing actions. However, in order to model crucial aspects of affordance-oriented processing, it is necessary to consider behaviours that require true physical interactions with objects, since this characteristic leads to an entirely different processing objective compared to classical perceptual processing, e. g. object recognition, where the physical environment is passively analysed. In this paper we will present first simulation results as well as an experimental verification of the model.

<sup>1</sup> School of Psychology, University of Birmingham Birmingham B15 2TT, United Kingdom, email: cxb632@bham.ac.uk

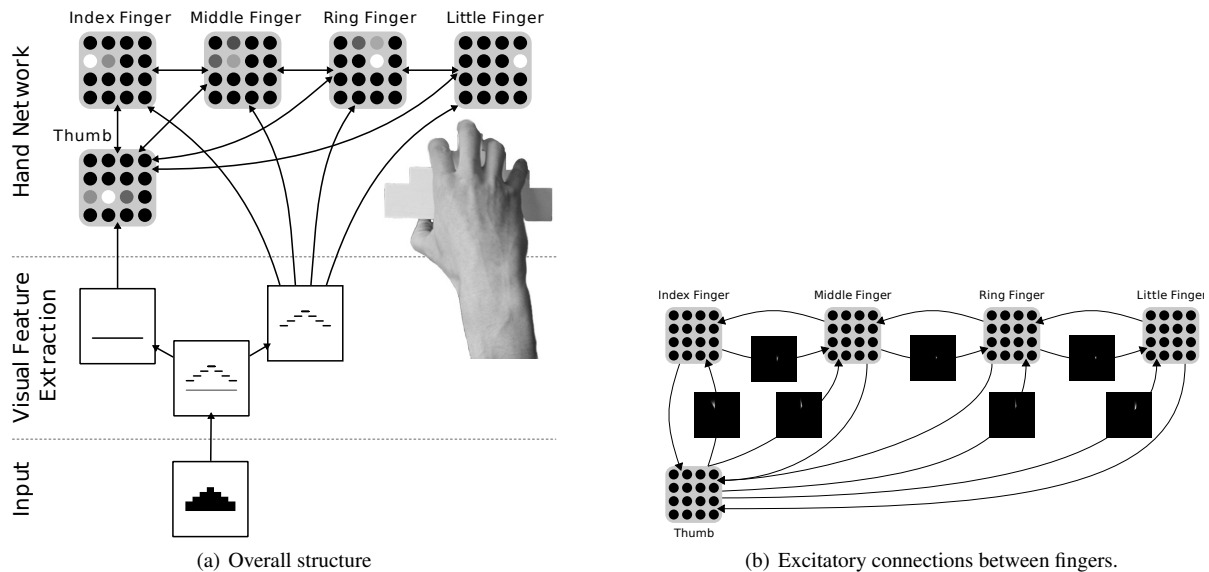


**Figure 1.** The Selective Attention for Action model (SAAM)

## 2 The Selective Attention for Action Model (SAAM)

Figure 1(a) gives an overview of SAAM. The input consists of black&white images. The output of the model is generated in five “finger maps” of a “hand network”. The finger maps encode the finger positions which are required for producing a stable grasp of the object in the input image. At the heart of SAAM’s operation is the assumption that stable grasps are generated by taking into account two types of constraints, the geometrical constraints imposed from the object shape and the anatomical constraints given by the hand. In order to ensure that the hand network satisfies these constraints we followed an approach suggested in Ref. [13]. In this soft-constraint satisfaction approach, constraints define activity patterns in the finger maps that are permissible and others that are not. Then we defined an energy function for which the minimal values are generated by just these permissible activity values. To find these minima, a gradient descent procedure is applied resulting in a differential equation system. The differential equation system defines the topology of a biologically plausible network. The mathematical details of this energy minimisation approach are given in the next section. Here, we focus on a qualitative description of the two types of constraints and their implementation.

The geometrical constraints are extracted from the shape of the object in the visual feature extraction stage. To begin with, obviously, only edges constitute suitable contact points for grasps. Furthermore, edges have to be perpendicular to the direction of the forces exerted by the fingers. Hence, only edges with a horizontal orientation make up good contact points, since we consider only a horizontal hand orientation in this first version of the model (see Fig. 1(a)). We implemented horizontal edge detectors using Sobel filters [14]. Finally, to exert a stable grasp, thumb and fingers need to be located at opposing sides of an object. This requirement was

realized by separating the output of the Sobel filters according to the direction of the gradient change at the edge. In fact, the algebraic sign of the response differs at the bottom of a 2D-shape compared to the top of a 2D-shape. Now, if one assumes the background colour to be white and the object colour to be black, the signs of the Sobel-filter responses indicate appropriate locations for the fingers and the thumb (see Fig. 1(a) for an illustration). The results of the separation feed into the corresponding finger maps providing the hand network with the geometrical constraints. Note that, of course, the assumptions about the object- and background-colours represent a strong simplification. On the other hand, this mechanism can be interpreted as mimicking the result of stereo vision. In such a resulting “depth image” real edges suitable for thumb or fingers could be easily identified.

The anatomical constraints implemented in the hand network take into account that the human hand cannot form every arbitrary finger configuration to perform grasps. For instance, the maximum grasp width is limited by the size of the hand and the arrangement of the fingers on the hand makes it impossible to place the index, middle, ring, and little finger in another order than this one. After applying the energy minimisation approach, these anatomical constraints are implemented by excitatory connections between the finger layers in the hand network (see Fig. 1(a) and 1(b)). Figure 1(b) also illustrates the weight matrices of the connections. Each weight matrix defines how every single neuron of one finger map projects onto another finger map. The direction of the projection is given by the arrows between the finger maps. For instance, neurons in the thumb map feed their activation along a narrow stretch into the index finger map, in fact, encoding possible grip sizes. Each neuron in the target map sums up all activation fed through the weight matrices. Note that all connections between the maps are bi-directional whereby the feedback path uses the transposed weight matrices of the

feedforward path. This is a direct result of the energy minimisation approach and ensures an overall consistency of the activity pattern in the hand network, since, for instance, the restriction in grip size between thumb and index finger applies in both directions. Finally, since a finger can be positioned at only one location, a winner-takes-all mechanism was implemented in all finger maps.

## 2.1 Mathematical Details

### 2.1.1 Visual Feature Extraction

The filter kernel  $K$  in the visual feature extraction process is a simple Sobel-filter [14]. In the response of the Sobel-filter the top edges of the object are marked with positive activation while the bottom edges are marked with negative activation. This characteristic of the filter is used to feed the correct input with the geometrical constraint applied into the finger maps and the thumb map. The finger maps receive the filter response with all negative activation set to zero. The thumb map, however, receives the negated filter response with all negative activation set to zero:

$$I_{ij}^{(f)} = \begin{cases} R_{ij} & \text{if } R_{ij} \geq 0, \\ 0 & \text{else.} \end{cases}$$

$$I_{ij}^{(t)} = \begin{cases} -R_{ij} & \text{if } -R_{ij} \geq 0, \\ 0 & \text{else.} \end{cases}$$

with  $R_{ij} = I_{ij} * K$  whereby  $I_{ij}$  is the input image.

### 2.1.2 Hand Network

We used an energy function approach to satisfy the anatomical and geometrical constraints of grasping. In Ref. [13] an approach is suggested where minima in the energy function are introduced as a network state in which the constraints are satisfied. In the following derivation of the energy function, parts of the whole function are introduced, and each part relates to a particular constraint. At the end, the sum of all parts leads to the complete energy function, satisfying all constraints.

The units  $y_{ij}^{(f)}$  of the hand network make up five fields. Each of these fields encodes the position of a finger.  $y_{ij}^{(1)}$  encodes the thumb,  $y_{ij}^{(2)}$  encodes the index finger, and so on to  $y_{ij}^{(5)}$  for the little finger. For the anatomical constraint of possible finger positions the energy function is based on the Hopfield associative memory approach [15]:

$$E(y_i) = - \sum_{\substack{ij \\ i \neq j}} T_{ij} \cdot y_i \cdot y_j.$$

The minimum of the function is determined by the matrix  $T_{ij}$ . For  $T_{ij}$ s greater than zero, the corresponding  $y_i$ s should either stay zero or become active in order to minimize the energy function. In the associative memory approach,  $T_{ij}$  is determined by a learning rule. Here, we chose the  $T_{ij}$  so that the hand network fulfils the anatomical constraints. These constraints are satisfied when units in the finger maps that encode finger positions of anatomically feasible postures are active at the same time. Hence, the  $T_{ij}$  for these units should

be greater than zero, and for all other units,  $T_{ij}$  should be less than or equal to zero. This lead to the following equation:

$$E_a(y_{ij}^{(g)}) = - \sum_{f=1}^5 \sum_{\substack{g=1 \\ g \neq f}}^5 \sum_{ij} \sum_{\substack{s=-L \\ s \neq 0}}^L \sum_{\substack{r=-L \\ r \neq 0}}^L T_{sr}^{(fg)} \cdot y_{ij}^{(g)} \cdot y_{i+s, j+r}^{(f)}.$$

In this equation  $T_{ij}^{(fg)}$  denotes the weight matrix from finger  $f$  to finger  $g$ .

A further constraint is the fact that each finger map should encode only one position. The implementation of this constraint is based on the energy function proposed in Ref. [16]:

$$E_{\text{WTA}}(y_i) = a \cdot \left( \sum_i y_i - 1 \right)^2 - \sum_i y_i \cdot I_i.$$

This energy function defines a winner-takes-all (WTA) behaviour, where  $I_i$  is the input and  $y_i$  is the output of each unit. This energy function is minimal when all  $y_i$  are zero except one, and when the corresponding input  $I_i$  has the maximal value of all inputs. Applied to the hand network where each finger map requires a WTA-behaviour, the first part of the equation turns into:

$$E_{\text{WTA}}^a(y_{ij}^{(f)}) = \sum_{f=1}^5 \left( \sum_{ij} y_{ij}^{(f)} - 1 \right)^2.$$

The input part of the original WTA-equation was modified to take the geometrical constraints into account:

$$E_t(y_{ij}^{(f)}) = - \sum_{f=2}^5 \sum_{ij} w_f \cdot y_{ij}^{(f)} \cdot I_{ij}^{(f)},$$

$$E_t(y_{ij}^{(1)}) = - \sum_{ij} w_1 \cdot y_{ij}^{(1)} \cdot I_{ij}^{(0)}.$$

These terms drive the finger maps towards choosing positions at the input object which are maximally convenient for a stable grasp. The  $w_f$  factors were introduced to compensate the effects of the different number of excitatory connections in each layer.

**The Complete Model** To consider all constraints, all energy functions need to be added, leading to the following complete energy function:

$$E(y_{ij}^{(f)}) = a_1 \cdot E_{\text{WTA}}^a(y_{ij}^{(f)}) + a_2 \cdot E_{\text{at}}(y_{ij}^{(f)}) + a_3 \cdot E_a(y_{ij}^{(f)}).$$

The parameters  $a_i$  weight the different constraints against each other. These parameters need to be chosen in a way that SAAM successfully selects contact points at objects in both conditions, single-object images and multiple-object images. The second condition is particularly important to demonstrate that SAAM can mimic affordance-based guidance of attention. Moreover, and importantly, SAAM has to mimic human-style contact points. Hereby, not only the parameters  $a_i$  are relevant, but also the weight matrices of the anatomical constraints strongly influence SAAM's behaviour.

**Gradient Descent** The energy function defines minima at certain values of  $y_i$ . To find these values, a gradient descent procedure can be used:

$$\tau \dot{x}_i = - \frac{\partial E(y_i)}{\partial y_i}.$$

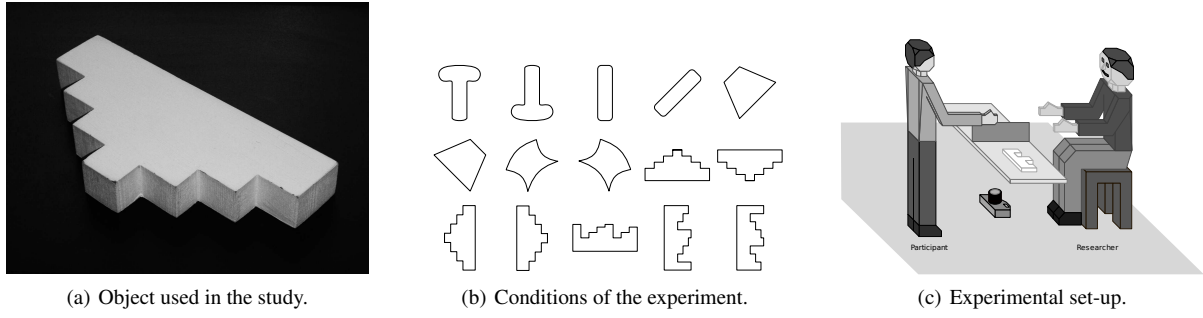


**Figure 2.** Material and procedure of the grasping experiment.

The factor  $\tau$  is antiproportional to the speed of descent.

In the Hopfield approach,  $x_i$  and  $y_i$  are linked together by the sigmoid function:

$$y_i = \frac{1}{1 + e^{-m \cdot (x_i - s)}},$$

and the energy function includes a leaky integrator, so that the descent turns into

$$\tau \dot{x}_i = -x_i - \frac{\partial E(y_i)}{\partial y_i}.$$

Using these two assertions, the gradient descent is performed in a dynamic, neural-like network, where  $y_i$  can be related to the output activity of neurons,  $x_i$  the internal activity, and  $\partial E(y_i)/\partial y_i$  gives the input to the neurons.

Applied to the energy function of SAAM, it leads to a dynamic unit (neuron) which forms the hand network:

$$\tau \dot{x}_{ij}^{(f)} = -x_{ij}^{(f)} - \frac{\partial E_{\text{total}}(y_{ij}^{(f)})}{\partial y_{ij}^{(f)}}.$$

To execute the gradient descent on a computer, a temporarily discrete version of the descent procedure was implemented.

### 3 Verification of the model

This study tested whether SAAM can generate expedient grasps in general and whether these grasps mimic human grasps. To accomplish this, simulations with single objects in the visual field were conducted. The results of the simulations were compared with experimental data on grasping these objects. In the following two sections we will at first present the experiment and its results and then compare its outcomes with the results from our simulations with SAAM.

#### 3.1 Experiment

We conducted an experiment in which humans grasped objects. Interestingly, there are only very few published studies on this question. Most notably D. P. Carey *et al.* examined grasps of a stroke patient [17]. However, no studies with healthy participants can be found in the literature.

**Participants** We tested 18 school students visiting the psychology department on an open day. The mean age was 17.8 years. All participants but two were right-handed. The left-handed participants were excluded from further analysis because the objects had not always been mirrored correctly during the experiment.

**Material** For the experiment we designed six two-dimensional object shapes. The objects were made of 2.2 cm thick wood and were painted white. Their size was between  $11.5 \times 4$  and  $17.5 \times 10$  centimetres (see Fig. 2(a) for an example). By presenting the objects in different orientations we created fifteen conditions (see Fig. 2(b)). Note that the shapes are highly unfamiliar, non-usable. Hence, the influence of high-level object knowledge is limited in the experiment. We chose this set-up in order to be compatible with the simulations in which SAAM possesses no high-level knowledge either.

**Procedure** Figure 2(c) illustrates the experimental set-up. During the experiment participants and experimenter were situated on opposite sides of a glass table facing each other. The glass table was divided in two halves by a 15 cm high barrier. Participants were asked to position themselves so that their right hand was directly in front of the right half of the glass table. In each trial the experimenter placed one of the objects with both hands in the right half of the glass table. The participants were then asked to grasp the object, lift it and place it into the left half without releasing the grip. The experimenter took a picture with a camera from below the glass table (see Figure 3(a) for an example). After taking the photo, the participants were asked to return the object to the experimenter. The last step was introduced to ensure that the participants would not release their grasp before the photo was taken. As soon as the object was handed back to the experimenter, a new trial started by placing the next object in the right half of the glass table. Each participant took part in two blocks with fifteen trials each. The order of the trials was randomised.

**Results** To analyse the pictures taken in the experiment, we developed a software for marking the positions of the fingers in relation to the objects. In Figure 3(b) the resulting finger positions are shown for the first condition. Even

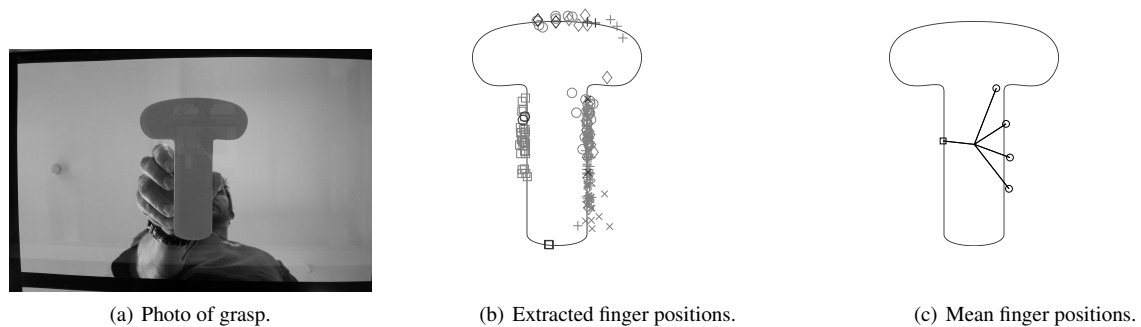


**Figure 3.** Mean finger positions: Finger positions (b) are extracted from photos (a). The fingers are colour-coded (thumb marked in red; right-hand grasp). For each finger its mean position is calculated (c). The thumb position is highlighted by a square box.

though the grasps show some variability, in general, participants grasped the object in two ways: they either placed their thumb at the left side of the object and the fingers on the right side or they placed the thumb at the bottom of the object and the fingers on the top edges. These two sets of different grasping positions are indicated with two markers in Figure 3(b) (circle and square). Such sets of different grasping positions were observed in all conditions.

To determine a “typical” grip from the experimental data, averaging across these very different set of grasping positions would not make sense. Therefore, we calculated the mean finger positions for each set of grasping positions separately. The resulting mean positions for the first condition are shown in Figure 3(c). Sets of grasping positions containing only one or two samples were discarded as outliers. For the comparison with the simulation results we only considered the set of grasping positions for each object chosen in the majority of trials.

### 3.2 Simulations

We conducted simulations with SAAM using the same objects as in the experiment. Figure 4 shows two examples of the simulation results. These illustrations also include the mean finger positions from the experimental results for a comparison with the simulation data. The ellipses around the mean finger positions illustrate the variations in the data. The comparison shows that most finger positions lie within the ellipses. Hence the theoretical assumptions behind SAAM that geometrical and anatomical constraints are sufficient to mimic human behaviour have been confirmed. Note that not all experimental conditions could be simulated with SAAM, since the model is currently only able to create horizontal grasps.

We also tested simulations with two objects in the visual field to test SAAM’s ability to simulate attentional processes. The simulations were successful in the sense that contact points for only one object were selected and the second object was ignored (see Conclusion for further discussions).

## 4 Conclusion and Outlook

Recent experimental evidence indicates that visual attention is not only guided by visual properties of visual stimuli but also by affordances of visual objects. This paper sets out to develop a model of such affordance-based guidance of selective attention. As a case in point we chose to model grasping of objects and termed the model the Selective Attention for Action Model (SAAM). To detect the parts of an object which afford a stable grasp, SAAM performs a soft-constraint satisfaction approach by means of a Hopfield-style energy minimisation. The constraints were derived from the geometrical properties of the input object and the anatomical properties of the human hand. In a comparison between simulation results and experimental data from human participants we could show that these constraints are sufficient to simulate human grasps. We also tested whether SAAM cannot only extract object affordances but also implements the guidance of attention through affordances by using two-object images. Indeed, SAAM was able to select one of two objects based on their affordance. The interesting aspect here is that SAAM’s performance is an emergent property from the interplay between the anatomical constraints. Especially, the competitive mechanism implemented in the finger maps is crucial for SAAM’s attentional behaviour. This mechanism already proved important in SAIM [3] for simulating attentional effects of human object recognition. However, it should be noted that SAAM does not select whole objects as SAIM does. But, since SAAM and SAIM use similar mechanisms, it is conceivable that they can be combined to form one model. In such a model SAIM’s selection mechanism of whole objects can be guided by the SAAM’s selection of contact points. Hence, this new model could integrate both mechanisms, selection by visual-properties and by action-related properties, forming a more complete model of selective attention.

Despite the successes reported here, this work is still in its early stages. First, we will need to verify the priorities of object selection predicted by SAAM. We also plan to include grasps with a rotated hand to simulate a broader range of experimental data. Finally, there is a large amount of experimental data on the interaction between action knowledge

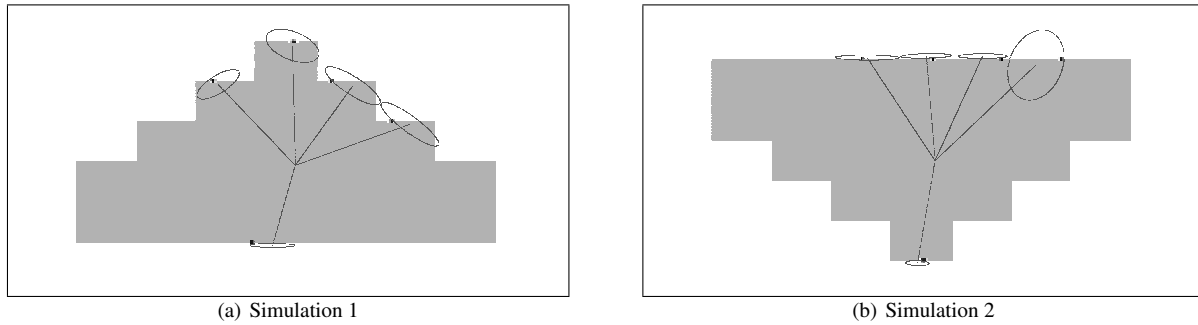


**Figure 4.** Comparison of experimental results and simulated grasps. The ellipses indicate the variation in the experimental data. The black dots mark the finger positions as generated by the simulations.

and attention (see Ref. [18] for a summary). Therefore, we aim to integrate action knowledge into SAAM, e. g. grasping a knife for cutting or stabbing. With these extensions SAAM will sufficiently contribute to the understanding of how humans determine object affordances and how these lead to a guidance of attention.

## REFERENCES

- [1] J. J. Gibson, *The senses considered as perceptual systems* (Houghton-Mifflin, Boston, 1966).
- [2] J. J. Gibson, *The ecological approach to visual perception* (Houghton-Mifflin, Boston, 1979).
- [3] D. Heinke and G. W. Humphreys, *Psychological Review* **110**, 29 (2003).
- [4] S. T. Grafton, L. Fadiga, M. A. Arbib and G. Rizzolatti, *NeuroImage* **6**, 231 (1997).
- [5] J. Grèzes and J. Decety, *Neuropsychologia* **40**, 212 (2002).
- [6] M. Tucker and R. Ellis, *Journal of Experimental Psychology* **24**, 830 (1998).
- [7] J. C. Phillips and R. Ward, *Visual Cognition* **9**, 540 (2002).
- [8] A. M. Borghi, C. Bonfiglioli, L. Lugli, P. Ricciardelli, S. Rubichi and R. Nicoletti, *Neuroscience Letters* **411**, 17 (2007).
- [9] T. C. Handy, S. T. Grafton, N. M. Shroff, S. Ketay and M. S. Gazzaniga, *Nature Neuroscience* **6**, 421 (2003).
- [10] G. di Pellegrino, R. Rafal and S. P. Tipper, *Current Biology* **15**, 1469 (2005).
- [11] A. H. Fagg and M. A. Arbib, *Neural Networks* **11**, 1277 (1998).
- [12] P. Cisek, *Philosophical Transactions of the Royal Society* **362**, 1585 (2007).
- [13] J. J. Hopfield and D. W. Tank, *Biological Cybernetics* **52**, 141 (1985).
- [14] R. C. Gonzalez and R. E. Woods, *Digital Image Processing* (Addison-Wesley, 1993).
- [15] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, in *Proc. of the Nat. Academy of Sciences*, (8)1982.
- [16] E. Mjolsness and C. Garrett, *Neural Networks* **3**, 651 (1990).
- [17] D. P. Carey, M. Harvey and A. D. Milner, *Neuropsychologia* **34**, 329 (1996).
- [18] G. W. Humphreys and M. J. Riddoch, *Psychology of learning and motivation* **42**, 225 (2003).

# The Logic of Robotics Inspired Biology

Martin Hülse and Mark Lee<sup>1</sup>

**Abstract.** Biologically inspired robotics is a well known approach for the design of autonomous intelligent robot systems. Very often it is assumed that biologically inspired models successfully implemented on robots offer new scientific knowledge for biology too. In other words, robots experiments serving as a replacement for the biological system under investigation are assumed to provide new scientific knowledge for biology. This article is a critical investigation of this assumption. We begin by clarifying what we mean by “new scientific knowledge.” Following Karl Popper’s work the *The Logic of Scientific Discovery* we conclude that in general robotic experiments serving as replacement for biological systems can never directly deliver any new scientific knowledge for biology. We further argue that there is no formal guideline which defines the level of “biological plausibility” for biologically inspired robot implementations. Therefore, there is no reason to prefer some kind of robotic setup before others. Any claimed relevance for biology, however, is only justified if results from robotic experiments are translated back into new models and hypotheses amenable to experimental tests within the domain of biology. This translation “back” into biology is very often missing and we will discuss popular robotics frameworks in the context of Brain Research, Cognitive Science and Developmental Robotics in order to highlight this issue. Nonetheless, such frameworks are valuable and important, like pure mathematics, because they might lead to new formalisms and methods which in future might be essential for gaining new scientific knowledge if applied in biology. No one can tell, if and which of the current robotics frameworks will provide these new scientific tools. What we can already say—the main message of this article—is that robot systems serving as a replacement for biological systems won’t be sufficient for the test of biological models, i.e. gaining new scientific knowledge in biology.

## 1 Introduction

Nature has been an inexhaustible and illuminating source of inspiration for scientists and engineers through all disciplines and it will continue to be so without any doubt. When engineering reaches its limits, nature might have good ideas how to proceed. In the area of autonomous and intelligent robot systems living beings outperform current engineering by far. Therefore not surprisingly a whole body of work in this field is entitled “biologically inspired robotics” where research is focused on copying biological systems. The hope is that this might lead to novel technologies which will help to close the current gap between natural and artificial intelligence. Whether or not this kind of biological inspiration has truly led to better robot systems shall not be the scope of this paper. We rather ask whether an implementation of a biologically inspired model on a robot sys-

tem can have any scientific value for biology. In other words, *can biologists gain new scientific knowledge from biological models implemented on robots?* This is the guiding question of this article.

The answer can have interesting implications. If the answer is positive then robotics or computer science in general could be seen as a discipline which provides scientific knowledge about biological phenomena. Hence, computer science would be part of the empirical sciences, which is in fact an “old dream” of Artificial Intelligence [24].

Once the far reaching consequences of this question has been acknowledged we must carefully proceed in our search for an answer. The objective of this article is to provide some solid ground which might also be a starting point for future and more comprehensive studies of this issue. We begin by explicitly defining what we generally mean by “gaining new scientific knowledge.” As the reader will see, the Philosophy of Science and Karl Popper’s work *The Logic of Scientific Discovery* [18] in particular will provide a coherent definition (Section 2). Applying this definition in the context of biology and robotics, we will come to the conclusion that robotic experiments can never directly deliver any scientific knowledge for biology (Section 3). However, there are frameworks claiming to be of “biological relevance.” We will critically discuss this issue in Section 3 where outline the difference between “new insights” and “new scientific knowledge”. We argue that the relevance of robotic systems for new scientific knowledge in biology can only lie in the potential future impact on biology by developing new formalisms, paradigms and scientific tools leading to advanced studies of biological phenomena. This is explained in Section 4 by concrete examples from Brain Research and Cognitive Science followed by a case study outlining our work in Developmental Robotics (Section 5).

## 2 Karl Popper’s Logic of Scientific Discovery

Since ancient times philosophers have been puzzled by the problem of how human beings are able to gain knowledge about the world if there is no logical justification that observations made in the past will occur in the future. Therefore, any theory or general law derived from past observations (inductive reasoning) is logically not valid. As a consequence, the truth of a theory inferred by inductive reasoning can not be proven, thus, the truth of such a theory is not ensured. This is called the problem of induction [22].

The problem of induction is addressed by Karl Popper in his main work *The Logic of Scientific Discovery* [18] where he rejects inductive reasoning as a method for theory-building. According to Karl Popper a logically valid foundation for knowledge discoveries in empirical sciences can only be provided by the process of falsification. The statements or hypotheses derived from a theory via deduction can be tested empirically by conducting experiments. If the predictions inferred from a hypothesis do not match with the outcome of

<sup>1</sup> Intelligent Robotics Group, Department of Computer Science, Aberystwyth University, SY23 3DB, Wales, UK, email: {msh, mhl}@aber.ac.uk

the experiment, then the theory must be refuted because it obviously leads to a wrong statement. And a theory which allows the deduction of wrong statements cannot provide a logically valid description of the phenomena under investigation. On the other hand, if the experimental outcome matches with the predictions of the theory then the theory is validated. Nonetheless, a validation does not tell us anything about the outcome of future experiments. Therefore, the truth of the theory remains questionable no matter how many times it has been validated, and moreover, validations not even make them more probable.

Karl Popper points out that scientific knowledge is gained through falsification only. Falsification is the only logically valid way that drives the development of new theories because falsification usually does not lead to the refutation of the theory as a whole. It rather guides a careful revision of it which will lead to a better theory. According to Popper, scientists therefore shall be aiming for experiments that falsify their theories rather than just validate them.

Notice, the origins of a theory, i.e. how human beings develop axioms or formal expression of general laws, is not the subject of Karl Popper's analysis. He even understands this issue as not relevant for the logical foundations of empirical sciences.

Without going into much detail, we have to clarify some terminology first before we can go on discussing Popper's view in the context of robotics and biology. Modern logic defines a theory as a set of sentences of a formal language. A model is an interpretation of the theory that makes all sentences of the theory true.

In empirical sciences models are understood to be interpretations of general laws. These laws are expressed in a form of calculus, like the equations of Newton's three laws of motion [10]. Taking the example from classical physics given in [10] one can say, general laws "are applied to a particular system, e.g. a pendulum, by choosing a special force function, making assumptions about the mass distribution of the pendulum etc. The resulting model then is an interpretation (or realization) of the general law."

Nowadays, the majority of discoveries in empirical sciences are presented as models. As long as a model is derived from an interpretation of a theory (a set of sentences of a formal language) then this model allows the deduction of hypotheses. Thus, models can be subject of a formal falsification process as Popper proposes it.

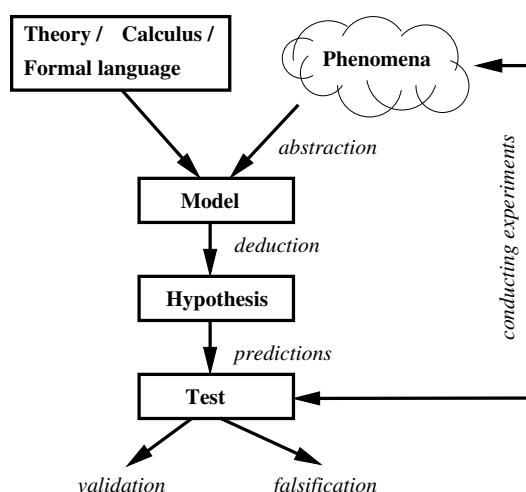


Figure 1. Schema of model-building and testing after Karl Popper [18]

Figure 1 presents a schema describing the process of testing models after Popper [18]. Interested in providing an explanation of some phenomena in a specific domain, scientists derive a model which is grounded in some formal framework. Every model comes with an abstraction of the target phenomena [20]. The level of abstraction is determined by the scientific question as well as by the formal framework (mathematical theory or axiomatic system). A hypothesis can be deduced from the model, which is a logically valid procedure. The derived hypothesis will provide predictions that might be amenable to experiments. If predictions of the model are confirmed by the experiments the model is verified. If not, the model is falsified and must be refuted as an explanation of the targeted phenomena. The latter outcome will lead to a revised and hopefully better model, in other words scientific knowledge will be gained by falsification.

It is important to note that in empirical sciences the *experimental tests of a hypothesis are conducted in the same domain as the phenomena under investigations*. Hypotheses about a biological system are tested by experiments with this biological system. Models of brain phenomena must be tested against data measured from real brains. If experiments are conducted not in the original domain of the target phenomena, then we cannot refer to such experiments as a test of the model for the original phenomena.

### 3 Robotics in Biology

#### 3.1 The problem of using robots for testing biological models

Regarding the usage of robot systems in biology we understand a robot system as an experimental device where the essential parts are assumed to be a realization or interpretation of a general law or a formal model. The aim of this realization is the experimental tests of a biological model. We assume models as a formal description which allows the deduction of hypotheses. Therefore, the robot system itself is not a model as often proclaimed. It is, including the computational processes running on it, part of an experimental setup. Similar to the example of the pendulum above, a robot is a device to generate data that helps to test the model it instantiates.

In addition we now stress the issue of the experimental domain. Recognising a robot as an experimental device, we have to ask in which domain the experimental data are generated. In other words, where does the data come from? It must be admitted that robot experiments serving as replacement of the biological system under investigation do not generate data in original domain of the biological phenomena. Hence, testing a hypothesis derived from a biological model, i.e. targeting biological phenomena, by exclusively using a robot system is not a logically valid test in the sense of Popper's Logic of Scientific Discovery. Experimenting with a robotic system introduces a new domain, that in general is totally different to the original biological domain of the target phenomena. Thus, any data derived from such robot experiments do not allow any conclusion about the explanational power of the model with respect to the biological phenomena. In summary, *robotic systems do not provide a logically valid test of a hypothesis about biological systems because the original biological domain is lost*. In consequence, from the point of gaining logically valid scientific knowledge in biology, the use of robot systems cannot be sufficient.

Notice, we do not argue that the change of domain is due to the use of technical equipment. Nowadays, probably any experiment in empirical sciences is in need of sophisticated technical apparatus in order to create controlled conditions and guarantee good measurements. Nevertheless, such devices still generate the requested data

out of the interaction with the target phenomena. This is not the case for a robot system that replaces the biological system. All the data generated by such a robot replacement are totally decoupled and isolated from the domain of the phenomenon a biological model is intended to address. This is what we call the loss of the original domain. Fig. 2 shows an attempt to illustrate this issue.

Leaving the domain B of biological phenomena when testing a hypothesis B derived from the original model has serious consequences. Applying Popper's framework, the predictions of hypothesis B turn into the target phenomena in a robotic domain R. This domain is grounded in a different formal theory which is determined by the computational paradigm, the technology and specific requirements. As a result a new model will be developed which might provide a similar level of abstraction and similar predictions but the test, the experiment, is done in a pure robotic domain. Hence, the loss of the original domain does not provide any conclusions about the explanatory power of model B or R for the phenomena in domain B.

### 3.2 Why robots at all?

The problem of losing the original domain when performing experiments with robots obviously questions the role of robot experiments in the context of biology. The usage of autonomous robots as scientific method in biology has been an object of debate since their very first appearance in the literature of brain research. Most famous is Grey Walter's book *The living brain* [27], where he introduces robots performing tropisms and even learning tasks. More recently the topic has been raised in articles by Barbara Webb [28, 29]. She proposes seven categories (called dimensions) that are intended to provide a guideline for the evaluation of the relevance of "robot models" for biological phenomena. In her 2009 article [29] she goes even further arguing that the only proof of biological relevance of "robot models" is the proposal of a new hypothesis that can be tested with the real biological systems. The latter statement is much in line with the conclusions in this work. However, in contrast to our argumentation she is not addressing the role of validation and falsification. This lack, it seems to us, might muddy the general demarcation between the

robotic domain and the world of biological phenomena. We deny any logically valid relations between biology and robot experiments, therefore, in general no robot experiment can provide logical valid falsification or validation of a model aimed at biological phenomena. In this respect our claims are much more strict because we say, independent of Webb's seven dimensions (Relevance, Level, Generality, Abstraction, Structural accuracy, Performance match, Medium [28]) models, hypotheses and experiments in the robot domain do not provide any new scientific knowledge for biology. Consequently, Webb's seven dimensions fail in their attempt to provide guidelines for what "good robot models" for biological phenomena are supposed to be.

Only the translation from robotics back into biological experiments allows validation or falsification, and thus, logical valid scientific knowledge. Thus any robot experiment can become a source of formal or informal inspiration that might be translated into the biological domain where it can be tested. Hence, what seems to be quite strict or limiting, i.e. no logical relation between robot domain and biology, provides in fact a much more relaxed view about the tools and methods used to "get inspiration" for good biological models. Since there is no logical relation between the domain of biology and robots, there is no need to introduce limitations in the robotic domain for the sake of its legitimation for biological phenomena. Whether it be a physical robot or a simulation, a detailed high-dimensional model of a specific brain structure or a very abstract and low-dimensional implementation of a dynamical system, they all have their value, if, in the short or the long run, models will be proposed which can be experimentally tested with the biological system under investigation. It is this point where our conclusion from Popper's Logic of Scientific Discovery opposes Barbara Webb's stand for the seven dimensions [28]. We argue, according to the *The Logic of Scientific Discovery* Webb's seven dimension cannot determine the biological relevance of a robot system.

In the following we briefly discuss common strategies used to justify or defend the biological relevance of specific robotic frameworks. Similar issues and many more are discussed in [29], but the reader will see that our proposal lets them appear in a different light.

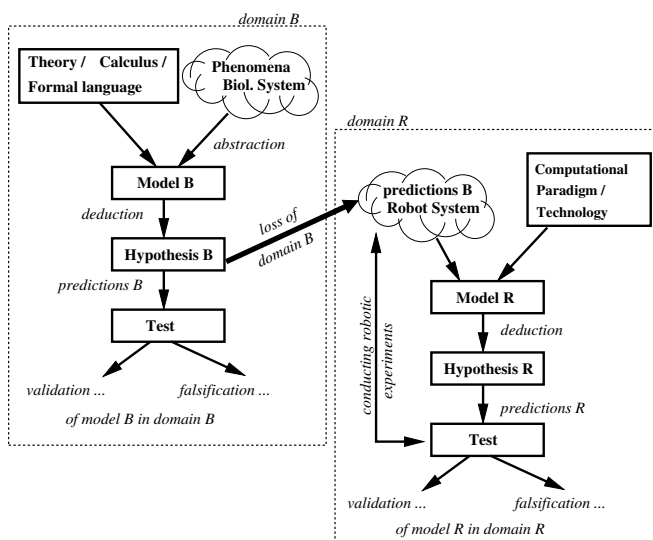


Figure 2. The loss of domain.

#### 3.2.1 Biologically inspired only

Biologically inspired robotics refers to robotic experiments where specific computational paradigms, novel algorithms or optimization processes copy mechanisms observed in biological systems. Sometimes it is claimed that the scientific value of the work is given in terms of a validation of a biological model [3].

According to Popper's framework, however, a validation does not provide any new scientific knowledge. Consequently and without even considering the loss of the biological domain, a successful robot implementation of a biological model does not provide any logically valid statement that this model is to be preferred to any other model targeting the same phenomena.

When presenting biologically inspired robot implementations only, in the best case this can only serve engineering by promoting more efficient methods for new tasks, domains or application; a scientific value for biology is not given.

An interesting question occurs when assuming a biological model confirming a hypothesis which is not confirmed when implemented as a robot model. Apart from the fact that we are not aware of any published work of this kind, the question needs to be addressed. Can biologically inspired robotics experiments serve as a valid falsification of the model for the biological phenomena? We argue that they

cannot. Assume a hypothesis derived from a biological model which is tested in the biological domain as well as it is tested in a setup where a robot replaces the biological system. Given that in the first case the hypothesis is validated while it is falsified in the robotic experiment, then the robotic experiment would hardly be accepted as a falsification of the model. The reason is obviously that a robotic implementation raises a lot of concerns about being a valid representation of the biological system. In other words, the loss of the domain does not allow robotics to be test bed for the falsification of models intended to address biological phenomena.

### 3.2.2 Proof of concept

A proof of concept, also called an existence proof, is usually introduced in order to overcome paradigms that are believed to be the only valid explanation of certain phenomena. As an example the general law: “There is no I without S” can be countered by showing: “There is an I without S”. The most famous example in Artificial Intelligence for this kind of counter example is Rodney Brooks’ article *Intelligence without representation* [5], where he targeted the dominance of the symbolic information processing hypothesis by demonstrating that advanced robot behaviours can be generated without the application of the symbolic information processing paradigm. However, Brooks never said: “All intelligence without symbolic representation.”

With respect to the relevance of existence proofs for an understanding of biological systems we have to recognise that a robotic experiment introduced as an existence proof in a specific robotic domain does not provide any more than the implementation of a biologically inspired model (see discussion above). Hence, in the best case an existence proof serves engineering but not empirical sciences.

### 3.2.3 Demonstrations of principles

W. G. Walter’s robot system were already mentioned. Interestingly, these robot systems were promoted as “part of a mirror of the brain” [27] because they demonstrate principles that must be considered when investigating real brains. In Cybernetics such a principle is the “feedback loop in which the environment is a component.” In this context, robots or machines are seen as a valid and illuminating “mimicry of life”.

Obviously, Walter presented his robots as devices for the validation of a biological model. As we argue this is questionable since the biological domain is lost. However, the issue of “demonstration” includes an additional quality. Demonstrating a principle in a very abstract form gives evidence about the involvement of mechanisms that might be an essential part of many phenomena in different domains. These mechanisms might therefore be worth studying in a formal way, i.e. mathematical framework for their own sake, without the need to refer back to the biological target phenomena.

Whether or not the results of these studies are in the short or the long term helpful to understand specific phenomena of biological systems, this can only be answered by “explicitly translating” the results back into models, hypotheses and experimental tests in the biological domain. Walter (and others) often pursued such model translation exercises as they were primarily interested in neuroscience.

### 3.2.4 Support the development of new theoretical frameworks, paradigms, and tools

The aspect of robots as demonstrators of principles is closely coupled with the development of new theoretical frameworks. Studying

principles, like feedback loops, in a purely formal framework bears the danger that formalisms are evolved which lose any match with real systems. Using robot systems can serve as a first kind of reality check. Demonstrating the outcome of new formalisms on a robot system gives *insights* about the constraints and valid boundary conditions of a formalism. In this respect we understand Webb’s statement that robot experiments provide insights “by working on real problems in real environment” [28]. Notice, we understand *new insights* as different from *new scientific knowledge*. Only if insights gained from robotics experiments are translated back into biological models and tested in the biological domain, they can become new scientific knowledge in biology.

### 3.2.5 Robots as tools for the development of new hypotheses amenable to experimental test

The best way to motivate a new formalism or theory for the old unsolved problems is the ability to infer new models which allow the deduction of already known and new predictions that can be tested experimentally. Real robots, simulations or even a set of equations are promoted as tools that help us to provide new testable hypotheses. Indeed, robots interacting within unconstrained environments or computer simulations often produce unforeseen but reproducible results. These unexpected experimental outcomes might be seen as an inevitable consequences of the model applied and therefore are claimed to be new predictions or hypotheses. Nevertheless, only the experimental test of these predictions in the biological domain allows conclusions whether or not “the new hypothesis” is just a superficial effect of the non-biological domain.

Moreover, the test of a new hypothesis for biological systems derived from experiments in non-biological domains involves an additional problem. If the “new hypothesis” cannot be derived from the biological model in a logically valid way—e.g. without computer simulations or experiments in the robotics domain—then the “new hypothesis” is not a logically valid deduction. Consequently, an experimental test of such a hypothesis makes no sense at all because one does not know whether this hypothesis is a logical consequence of the biological model.

## 4 Current robotic frameworks in brain research and cognitive science

In the following we examine popular robotic frameworks in brain research and cognitive science where we briefly discuss how they match with our five “categories of justification of biological relevance.”

### 4.1 Brain Based Devices

Following specific design principles a brain-based device, introduced and promoted by G. M. Edelman [9], is a physical robot system interacting autonomously with its environment. The control structure of the system is provided by “simulated nervous systems”, in fact large-scale artificial neural networks. In order to allow comparison with real brain-data the design of the artificial neural networks tries to “reflect the brain’s anatomy and physiology” [14]. A series of robot experiments demonstrating a variety of adaptation tasks driven by neural group selection principles is for instance given in [14] where it is stated that brain-based devices serve as “a heuristic for testing theories of brain function.” With respect to “testing” it must be admitted that although the hypotheses and predictions are formulated

in the terminology of brain science they are experimentally tested with robot systems only. In this sense the experiments presented have primarily the character of a validation of a robotic model.

Regarding the validation of a biological model it is claimed that the results of the robot experiments are in accordance with brain data from literature. These data, however, were published over ten years earlier. Thus, it is hard to see how the introduced robotic experiments have led to a new hypothesis.

Although strictly designed in relation to a specific level of abstraction in brain science, the experiments of the brain-based devices remain in the robotics domain. Moreover, experiments have the character of validation. It is for instance shown that the “binding problem” can be solved in a particular robot-environment system by synchronisation mechanisms. But as long as these results are not transferred into models and new testable hypotheses for real brains, the scientific value for brain research remains questionable.

From an engineering perspective brain-based devices might offer alternatives to symbolic information processing strategies for the development of advanced robot systems. As the authors emphasise brain-based devices “provide groundwork for the development of intelligent machines that follow neurobiological rather than computational principles in their construction” [14].

In conclusion brain-based devices can be seen as tools for the validation of biologically inspired models of neural information processing and the demonstration of principles (like synchronisation). An explicit formulation of biological models leading to new hypotheses testable in real brains has not been shown yet.

## 4.2 Conceptual neural evolution

Understanding the complexity of the human brain by investigating how this complexity is rooted in “simpler complexities” of “our evolutionary cousins” is proposed by M. A. Arbib as the framework of conceptual neural evolution. This approach emphasises the principle of “evolutionary refinement” where new behavioural competences emerge from the modulation of already given but simpler behaviours [1]. Arbib states explicitly that the models developed in this framework are derived from and tested with data from real animals. Hence, biologically inspired robot systems are a side effect of this research which “will advance the design principles for a new generation of biologically inspired robots” [1]. One example, a computational model for path planning [2], is explicitly mentioned as a source of inspiration for behaviour-based robotics [1]. Hence, within conceptual neural evolution Arbib sees robot systems as demonstrators of principles of “evolutionary refinement” only. As a source of inspiration for the development of brain models robotics plays no part.

## 4.3 Evolving Minimal Cognitive Agents

Within the context of cognitive science the application of evolutionary algorithms to generate robot behaviour for real or simulated robot systems is motivated as the evolution of “minimally cognitive behaviour” [4]. Such experiments are meant to be a step towards abstracting general principles of cognition that can be explicitly expressed within the formalism of Dynamical Systems Theory [26]. Since the process of artificial evolution allows the generation of low-dimensional neural circuits, the resulting control structures can be analytically investigated in a reasonable depth.

Scientists in this area, like Randal Beer, admit that the evolution of minimally cognitive behaviour is a result of an abstraction process

which allows no comparison with or empirical predictions for biological phenomena. Thus, the main contribution of this framework is to provide concrete examples within the specific formalisms of Dynamical Systems Theory, because “only by working through the details of tractable concrete examples of the cognitive phenomena of interest that we can begin to glimpse whatever general theoretical principles may exist and whatever mathematical and computational tools will be necessary to formulate them” [4].

Consequently, this approach fits into the category which we have introduced above as the “support of the development of new theoretical frameworks, paradigms and tools.” Once it is explicitly stated that the “evolution of minimally cognitive behaviour” is totally decoupled from the biological phenomena, we are comfortable with following Beer’s argumentation about the possible biological relevance of such studies in the near or far future.

## 5 A case study from Developmental Robotics

Among psychologists there is a longstanding controversy regarding the origins of many brain functions and observed behaviours. On one hand, the nativists argue that much basic function is innate and is thus specified genetically and becomes hardwired during neural maturation. On the other side, the empiricists claim that much apparently innate competence is instead formed through experience or interaction with the environment and hence involves some kind of adaptation or learning. Much of this debate centres on infant development, particularly the early growth of sensory-motor skills and cognition. Some behaviours, such as the early reflexes, are clearly innate because they have no opportunity for learning and are immediately available. Other functions, such as smiling and responding to faces, are less easily analysed and there are arguments on both sides for their emergence and growth. We now discuss a model of eye movement that illustrates how robotic models may shed some light on the general principles involved in one small aspect of this general problem. It is important to note that our model is not able to comment on the biological structure and validate or falsify particular features, but it is possible to expose some principles of logical necessity that must be present (or absent) in both the robot model and the biological system.

The human newborn is helpless at birth but has available a range of automatic sensory-motor reflex behaviours and also some primitive but potentially fundamental competences that are precursors for cognitive growth. These include basic head orientation to audio stimuli [16] and saccading with the eyes to look at a visual stimulus [21]. Because newborns are able to saccade immediately (they do not have smooth pursuit eye control) this has often been taken as evidence that this is an innate function. We have built a robot system that very quickly learns how to saccade and this demonstrates the possibility that the fundamental process of visual saccading to a peripheral stimulus could be learned rather than innate. If such rapid learning does occur it would be quite difficult to detect but it could be very significant evidence for the empiricist stance [25].

In computer vision or robotics, if a camera (robot “eye”) is to be moved in order to bring an object on the periphery of the image towards the centre then this is usually achieved by complex calculations involving knowledge of the geometry of the camera location, the scene structure and the parameters of the optical setup [30]. Even in systems that try to minimise the engineering of this task there is the need to calibrate the geometric and optical configuration [8].

As part of our work in Developmental Robotics we have built a model that does not need any prior information and is self-

calibrating. Through a process of learning the model essentially performs a kind of continuous calibration in which the mapping between image movement and eye movement is learned. The details are described elsewhere [7] but the method is straightforward. Our robot eye is implemented as two sensory functions: image processing software is used to simulate a periphery sensor and a centre or foveal sensor. The periphery sensor detects the position of any phasic changes in the visual periphery area. The centre sensor analyses any objects (i.e. colour blobs) in the foveal region of the visual field. This configuration is similar to the human vision system which drives saccades towards peripheral stimuli in order to fixate the foveal processing machinery onto targets that attract attention. The main control issue for this ocular-motor system is a sensory-motor coordination problem: what are the necessary motor variables to drive the eye so that a specific sensed region in the image is moved to the centre of the image? This means a transform must be known for converting image space displacement vectors into eye gaze space movement vectors. We designed a simple algorithm that can learn this transform as a mapping between the spaces. Local spatial correspondences are established on the basis of experience and are entered into the mapping in a continuous, cumulative and incremental process. The underlying assumptions are that (a) the mapping may not be a simple or linear relationship and (b) the position of the eyeball can not be related to image data until after birth (as vision is ineffective in the womb [23]) and therefore learning should start from zero prior knowledge.

If we start with no prior information then when the first peripheral target appears there will be no appropriate motor values known and some spontaneous movement must be made in order to gain some experience. When eventually the eye finds the foveal region then the parameters can be recorded in the mapping (initial peripheral location and final angle of gaze) for future reference. After a while sufficient regions become mapped so that only a few movements are needed before finding the target, and eventually the map becomes fully populated so that any target can be brought to the centre in a one step saccade. We observed three distinct qualitatively different behaviours emerge during this learning process: at first random Brownian-like wandering appeared, followed by more directed multi-step paths that appear to involve corrections, and finally single accurate saccades straight to the centre. These behavioural patterns emerged successively although not in completely strict sequence.

The most significant results from this experiment are: the system learns very rapidly; it does not require any calibration process or prior knowledge; it continuously adapts to correct errors; and displays qualitatively different stages in its behaviour which emerge during learning. Perhaps the most important result is the extremely fast learning speed – which was much faster than current neural network based approaches. The simplicity of the method is important in this regard, as the requirement for fast performance, in both saccading and learning, rules out complex computations and the speed of neural processes limits the number of steps that can be involved [17]. We found that some 200 saccade attempts were sufficient to build a near complete mapping, and if this is converted into infant activity the figures show that a saccade map might be built in a matter of hours, or at the most a few days.

Other results from our model show compatibility with the known data on early infant eye activity. Specific infant data on the size, eccentricity and average number of saccades, corroborate the behaviours observed [21, 11]. The accuracy of saccades with our model is a good match with reported infant accuracy [15, 13]. The average error in saccading to a given image location is 0.35 of the tolerance spacing between points in the mapping.

To return to the role of such models, we now see that it is *logically possible* that newborn saccadic eye behaviour could be learned rather than innate. This learning could be so rapid that it might not be easy to detect and the onus is now on psychology and neuroscience to investigate for new evidence. Others have reported similar fast learning in infants across a range of activities [19] and it has been suggested that the preference of newborns to orient towards faces could be learned very rapidly, even in the first six minutes of life [6]. Face orientation is another behaviour that is generally believed to be innate.

With this new viewpoint we can now reflect that it is quite difficult to see how accurate saccades could be created pre-wired before birth. The concrete and specific information needed to coordinate eyeball movement with movement on the image (i.e. movement *of* the image with movement *within* the image) depends upon the physical and optical properties of each particular eye and individual differences in anatomy and geometry. While the proprioceptive feedback from eyeball muscles can be calibrated in the womb, there are no detailed visual signals available prenatally for similar calibration of the image space [23]. And so it seems inconceivable that all the necessary information could be pre-compiled before birth.

Of course, just because we have similarity in results we do not claim that our, rather abstract, robot mechanism is the same or even similar to that which drives the human eye; there can be very many mechanisms that could produce the same behaviour. But what we have shown is that it is possible to learn the necessary information for driving eye saccades very rapidly, and this general principle undermines the nativist assumption that newborn saccading is innate. This means the nativists are obliged to revisit their assumptions and provide better evidence for their stance. The new hypothesis that newborn infants may perform, as yet undetected, rapid learning also offers motivation for biological and psychological investigation to find evidence to either support or refute the idea.

Another aspect concerns the difficulty of observing learning behaviour. Because saccades are such fast movements there is no time for (neural) feedback, and therefore the control must be a forward model [12]. Harris has argued that in this situation “motor babbling” is not just random “neural noise” but is “spontaneous information-gathering action”. We notice that the accurate, learned saccades are often mixed in with more exploratory movements (as gaps in the map are found and filled in). This makes the “motor babbling” much less obvious and it will require careful investigation to discern such learning from behavioural data alone.

Finally, another implication from our model shows the value of proposed mechanisms that may transfer between domains. It is well known that, for saccade errors, undershoot is more common than overshoots. An hypothesis has been proposed [12] that this is an optimum strategy to minimise the total flight-time, because the total flight-time is less with corrective saccades that undershoot as compared with those that overshoot. This effect also occurs in our method: fields near the fovea are predominant among the first to develop because most moves end in such a field, and so when a neighbour is selected it is more likely to be on the near side than the far side of the target. We analysed the data for a run of fixations and found that undershoot occurred in 75% of the cases. Further investigation, with both robotic and biological models may shed some light on the nature of such error patterns and the underlying mechanisms.

## 6 Conclusion

Facing the problem of induction in empirical sciences we followed Karl Popper's *Logic of Scientific Discovery* where a clear distinction is made between experimental validation and falsification and their role in gaining new and logically valid scientific knowledge. We have outlined how Popper's Logic of Scientific Discovery helps us to understand the fundamental problems when robotic experiments are motivated as a test beds for biological models.

The first of our conclusions was that a robot (simulated or real) is not a model. It is an experimental device to test models empirically. Moreover, when asking what kind of models can be tested with robot systems, we had to recognize the loss of the biological domain. The problem of losing the original biological domain means that, under any circumstances, experiments within a robotic domain cannot count as falsification or validation of biological models targeting biological phenomena. Thus, robotics does not directly provide new scientific knowledge for biology. As a consequence robot experiments can only be seen as an inspiration for new biological models and hypotheses. However, we only can accept the robot experiments as biologically relevant if they are explicitly translated back into the biological domain leading to new models and experimental tests. It is exactly this last step which is usually missing when robot experiments are motivated as biologically relevant.

In addition we must admit that there is obviously no formal way for transforming the results from robotics experiments into a biological model and corresponding experimental tests. Thus, robotics will always have a hard time to underline its scientific contribution to biology. In contrast to Barbara Webb we are not interested in tackling this problem by providing guidelines for "good robot models" that lead to "good biological models" [28]. These guidelines can only blur the general demarcation between robotics and biology. This demarcation can hardly be overemphasised since it forces scientists to translate results from robotics experiments back into new biological models and hypothesis amenable for tests in the original biological domain before they can claim that robots experiments are essential part of gaining new scientific knowledge in biology.

In the absence of clearly stated examples of "robotics inspired biology" we have asked for other qualities that provide indirect scientific value for biology. In addition to the obvious quality "Developing new hypotheses and experimental tests", we have stated the following four:

- Biologically inspired implementations
- Proof of concept
- Demonstrations of general principles
- Support of the development of new formalisms, paradigms, and tools

We have seen that any robot implementation of biologically inspired models can only count as a contribution to engineering but not to biology. The very same holds for the so called "existence proofs" or "proofs of concept" in robotics.

The two other qualities serve the development of new formalisms which are believed to provide the more powerful tools for the phenomena not yet understood. However, as long as these tools are not developed yet in such a way that they lead to new hypotheses and experiments they are of course subject of debate and must be carefully motivated [29].

We have discussed four different research activities or programs, namely

- Brain-based devices,
- Conceptual neural evolution,
- Evolving minimal cognitive agents, and
- Developmental robotics.

All these examples have shown that our criticism about the missing "direct biological relevance" is valid. Brain-based devices in particular are motivated as tools that help us to understand how the human brain works. The published results, however, stop at a level where it is claimed that a specific results from the robotic experiments suggest that a certain mechanism "might also be a common property of the neural networks of living animals" [9]. But, this remains an open question no matter how many robotic experiments are performed. Hence, according to our argument, biological relevance of brain-based devices remains open as well.

On the other hand, we have seen that the framework of artificial evolution of minimally cognitive agents is an example for the development of very abstract formalisms without even trying to generate new testable predictions for biological phenomena. As for pure mathematics, no one can tell yet whether this approach will lead to useful tools helping us to gain deeper understanding of biological phenomena. Only time can tell.

It is interesting to notice Arbib's strict view that the models in the framework of "conceptual neural evolution" are not inspired by any approach in robotics, but "will advance the design principles" for intelligent robots [1]. In this sense one could get the impression that Arbib primarily motivates his approach to Brain Research by the current limitations in engineering.

Finally, the extended case study of the learning of a robotic ocular-motor coordination has illustrated the demonstration of two principles. Firstly, the self-calibration of eye saccade movements is necessary and secondly, fast self-calibration is possible. These results gave evidence that saccadic eye behaviour of newborns could be learned rather than innate. Furthermore, with respect to the characteristic saccade errors observed in biological systems a similar dominance of undershooting saccades was found for in our robotic setup. Hence, the dominance of undershoots could be caused by mechanisms present in both systems, human eye and robot. The robotic system allows a formal description of the underlying processes leading to a comprehensive explanation of the characteristic dominance of undershoots. This in turn might help to derive a model explaining this phenomena for biological systems and leading to predictions which are amenable to experimental tests of the biological system. However, as long as this last step (models amenable to experimental test) is missing, we cannot claim biological relevance for our robotics experiments even if the resulting behaviours show similar characteristics.

At this point we conclude this article by stating that computer science or robotics do not belong to empirical sciences because we have shown that in general a robotic experiment can not provide new scientific knowledge about a biological phenomenon. We hope the reader will understand this as good news, since there is no need to limit the design of intelligent information processing systems by specific paradigms or levels of abstraction dictated by biology. The value of robotics and computer science for biology can only become manifested in the demonstration of general principles or the development of new formalisms, paradigms and/or tools. The assumption that biologically motivated robotic implementations directly provide new scientific knowledge for biological phenomena is logically not valid.

## ACKNOWLEDGEMENTS

The authors would like to thank Alison Pease and the anonymous reviewers for their valuable and constructive comments. Furthermore, we acknowledge the support by the EU-FP7 projects IM-CLeVeR (ICT-231722) and ROSSI (ICT-216125), and by EPSRC, UK, through grant EP/C516303/1.

## REFERENCES

- [1] M. A. Arbib, 'Rana computatrix to human language: towards a computational neuroethology of language evolution', *Phil. Trans. R. Soc. Lond. A*, **361**, 2345–2379, (2003).
- [2] M. A. Arbib and D. H. House, *Depth and detours: an essay on visually guided behavior*, 129–163, Vision, brain and cooperative computation (ed. M. A. Arbib and A. R. Hanson), Cambridge, MA: MIT Press, 1987.
- [3] M. A. Arbib, G. Metta, and P. van der Smagt, *Neurobotics: From Vision to Action*, 1453–1480, Springer Handbook of Robotics (ed. Siciliano and Khatib), Springer, 2008.
- [4] R. Beer, 'The dynamics of active categorical perception in an evolved model agent', *Adaptive Behavior*, **11**(4), 209–243, (2003).
- [5] R. Brooks, 'Intelligence without representation', *Artificial Intelligence*, **47**, 139–159, (1991).
- [6] N. J. Butko, I. R. Fasel, and J. R. Movellan, *Learning about humans during the first 6 minutes of life*, In: Fifth International Conference on Development and Learning, Bloomington, Indiana, USA., 2006.
- [7] F. Chao, M. H. Lee, and J. J. Lee, 'A developmental algorithm for ocular-motor coordination', *Robotics and Autonomous Systems*, DOI information: **10.1016/j.robot.2009.08.002**, (2010).
- [8] H. Chen, K. Matsumoto, J. Ota, and T. Arai, 'Self-calibration of environmental camera for mobile robot navigation', *Robotics and Autonomous Systems*, **55**(3), 177–190, (2007).
- [9] G. M. Edelman, 'Learning in and from brain-based devices', *Science*, **318**(5853), 1103–1105, (2007).
- [10] R. Frigg and S. Hartmann, *Models in Science*, E. N. Zalta (Ed.) Stanford Encyclopedia of Philosophy, 2006.
- [11] L. Hainline, J. Turkel, I. Abramov, E. Lemerise, and C.M. Harris, 'Characteristics of saccades in human infants', *Vision Research*, **24**(12), 1771–1780, (1984).
- [12] C. M. Harris, 'On the optimal control of behaviour: a stochastic perspective', *J Neurosci Methods*, **83**(1), 73–88, (1998).
- [13] C. M. Harris, M. Jacobs, F. Shawkat, and D. Taylor, 'The development of saccadic accuracy in the first seven months', *Clinical vision sciences*, **8**(1), 8596, (1993).
- [14] J. L. Kirchmar and G. M. Edelman, 'Brain-based devices for the study of nervous systems and the development of intelligent machines', *Artificial Life*, **11**, 63–77, (2005).
- [15] E. Kowler and E. Blaser, 'The accuracy and precision of saccades to small and large targets', *Vision Res.*, **35**(12), 1741–1754, (1995).
- [16] D. Muir and J. Field, 'Newborn infants orient to sounds', *Child Development*, **50**(2), 431–436, (1979).
- [17] N. Nagarajan and C. Stevens, 'How does the speed of thought compare for brains and digital computers?', *Current Biology*, **18** (17), 756758, (2008).
- [18] K. Popper, *Logic of Scientific Discovery*, Routledge London, New York, Baltimore, reprint 2009 edn., 2009.
- [19] P. Rochat and T. Striano, 'Emerging self-exploration by 2-month-old infants', *Developmental Science*, **2** (2), 206218, (1999).
- [20] A. Rosenblueth and N. Wiener, 'The role of models in science', *Philo. Sci.*, **12**, 316–321, (1945).
- [21] A. Roucoux, C. Culee, and M. Roucoux, 'Development of fixation and pursuit eye movements in human infants', *Behaviour Brain Research*, **10**, 133–139, (1983).
- [22] B. Russell, *The Problems of Philosophy*, Oxford Paperbacks, 2nd edn., 2001.
- [23] B. Schaal, B. Hopkins, and S. Johnson, 'Prenatal development of post-natal functions', (2005).
- [24] H. Simon, 'Artificial intelligence: An empirical science', *Artificial Intelligence*, **77**, 95–127, (1995).
- [25] J. Triesch, C. Teuscher, G. Deak, and E. Carlson, 'Gaze following: why (not) learn it?', *Developmental Science*, **9** (2), 125147, (2006).
- [26] T. van Gelder, 'The dynamical hypothesis in cognitive science', *BEHAVIORAL AND BRAIN SCIENCES*, **21**(5), 615–628, (1998).
- [27] W. G. Walter, *The Living Brain*, Harmondsworth, Middlesex: Penguin Books, 1961.
- [28] B. Webb, 'Can robots make good models of biological behaviour?', *BEHAVIORAL AND BRAIN SCIENCES*, **24**(6), 1033–1050, (2001).
- [29] B. Webb, 'Animals versus animats: Or why not model the real iguana?', *Adaptive Behavior*, **17**, 269–286, (2009).
- [30] Z. Zhang, 'A flexible new technique for camera calibration', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **22** (11), 1330–1334, (2000).

# AI, Attachment Theory and Simulating Secure Base Behaviour: Dr. Bowlby meet the Reverend Bayes.

Dean Petters<sup>1</sup>, Everett Waters<sup>2</sup>

**Abstract.** In this paper we describe how information processing constructs originating in AI have become part of the Attachment Theory tool kit. We survey the early influence of AI within the theoretical framework that John Bowlby formed as the foundation of Attachment Theory between the 1950s and 1980s. We then review recent work which has built upon Bowlby's framework and is concerned with modelling and simulating attachment phenomena. We conclude by discussing some possible advantages that might arise from incorporating recent work on Bayesian arbitration into attachment control system models.

## 1 Introduction

John Bowlby formulated his Attachment Theory whilst working in a multidisciplinary team that included childcare professionals, psychoanalysts, ethologists and other researchers at the Tavistock Clinic for over thirty years after the Second World War (10; 21). Bowlby was interested in understanding issues such as: the separation distress exhibited by children when they or their mothers were absent due to the infant's or mother's hospitalization (9); the effect of early maternal deprivation on later development (4); and grief and mourning in infancy (5). From its inception Attachment Theory integrated concepts from academic fields as diverse as Ethology, AI and Psychoanalysis. One of Bowlby's primary goals was to replace Freud's drive theory with an attachment motivation theory rooted in modern scientific methods, empirically accessible, and better able to account for infant behavior's sensitivity to context. To accomplish this, he first turned to ethology and developed a framework that described the attachment system as an instinctive behaviour system (29). However, in his more mature theoretical work, Bowlby drew increasingly on control systems theory and on AI based representational constructs such as Internal Working Models and hierarchical planning.

Bowlby's Attachment Trilogy (6; 7; 8) contains his most developed behavioural descriptions and theoretical explanations for attachment phenomena. These accounts comprise a set of observable behaviours related to social and emotional attachment in animals and humans; and the cognitive mechanisms that give rise to these observable phenomena. Attachment Theory can therefore be presented as incorporating two theoretical components that can be termed the behavioural and cognitive components (21). Attachment Theory's behavioural component is valuable for researchers interested in modelling and simulation because the set of attachment behaviours that have been observed and recorded over the last half century provide data for nu-

merous behavioural scenarios. These behavioural descriptions can be interpreted and adapted to act as standardised specifications of requirements allowing simulations to be focused in their evaluation (25; 26). Bowlby did not describe Attachment Theory's cognitive component in enough detail to be straightforwardly implemented in simulation without additional interpretation and development. What this can provide is a point of departure for contemporary AI theories. This means that simulation designers can use the information processing explanations for attachment phenomena that already exist within Attachment Theory as a starting point and inspiration when incorporating new AI structures and mechanisms.

## 2 The Behavioural Component of Bowlby's Attachment Theory

The offspring of many animals, including humans, show a tendency to gain and retain physical proximity to their main carer, usually their mothers. This is not surprising, as for many of these species the mother feeds her offspring, and infant animals have to approach to be fed. Some additional proximity seeking by the offspring to its mother may be due to secondary reinforcement effects that are driven by the primary reinforcement of gaining food. However, animal studies have shown proximity seeking to a main carer can be unrelated to any reinforcing effect of feeding. Research on imprinting in Geese (23) and research with infant monkeys and wire-frame mother substitutes (20) are amongst a number of studies that show animals can imprint or attach to individuals or objects that do not provide a source of food.

From the late 1930s Bowlby was undertaking human studies on attachment that can be seen as a parallel to research on attachment processes in animals. The key insight in interpreting the attachment behaviours linked to proximity seeking that Bowlby described is to see the attached infant or child as using their carer as a secure base from which to explore. When infants develop the ability to crawl, and later to walk, they can explore the world more effectively and the rate at which they acquire knowledge about the broader world accelerates. However, this new found ability to explore brings with it the potential for access to many more hazards. A naturalistic study of toddlers between ages 1 and 2 investigated how infants balance the opportunity for exploration and the security provided by their carers (2). In London parks, infant and carer pairs were observed without their awareness, in observation periods averaging 15 minutes' duration. This study found that most infants moved away from their carers to explore, but kept within a caregiver's line of sight, and periodically 'checked-in' by gaining the attention of their carers or by moving back to closer proximity.

Attachment development and individual differences have also

<sup>1</sup> Newman University College, Birmingham, UK, email: d.d.petters@cs.bham.ac.uk

<sup>2</sup> State University of New York at Stony Brook, email: everett.waters@sunysb.edu

been examined through the lens of a standardised laboratory procedure - the Strange Situation (1). This procedure is not an experiment where subjects are randomly assigned to different conditions in the laboratory. Rather, it presents all infants with the same controlled and replicable set of experiences. Nested within the normative trends that illustrate infant's sensitivity to context are several patterns of response reflecting the infant's confidence in the caregiver's response patterns.

A key research goal for the Strange Situation procedure was to demonstrate experimentally that infant attachment behaviour was sensitive to context - as opposed to rising and falling with the drive states of psychoanalytic theory (which would have a longer and more regular course and no obvious link to context). To capture infant responses to changes in context, the Strange Situation procedure consists of 8 three minutes episodes which are designed to activate, intensify or relax one-year-old's attachment behaviour in a moderate and controlled manner. The context changes that occur in the transitions between the eight episodes, and the and the infant's responses to these transitions, don't just help test a control systems approach against a drive state approach to explaining attachment. They also provide a valuable data-set for contemporary researchers interested in designing attachment behaviour simulations.

In the Strange Situation, the infant and carer enter the laboratory setting together, but then undergo a separation, when the carer leaves from the room, before a reunion in a subsequent episode. After the first reunion episode the infant also meets an unfamiliar 'stranger' in the laboratory, before a further separation. In each episode the infant's behaviour is carefully recorded from behind a two-way mirror. In the final episode the mother returns to her one-year-old infant after the infant has been left alone for three minutes in the unfamiliar setting. The infant's response in the reunion episodes correlates strongly with patterns of maternal behaviour and infant responses intensively observed throughout the previous year. An infant's responses to reunion in the Strange Situation can therefore act as a shorthand for the infant's home relationship with their carer (1). Normative behaviour patterns across episodes highlighted the infants' sensitivity to context. There were also marked individual differences which clustered into three distinct patterns, labelled: Secure (type B), Avoidant (type A), and Ambivalent (type C). More recent studies have categorised a fourth type of Disorganised pattern of attachment (type D) that is the least well characterised or understood and forms a very small proportion of infants in the general population (19, page 26).

Secure infants are the largest group and secure behaviour is the reference pattern against which the other classifications are evaluated. These infants respond to their mothers on reunion in the Strange Situation by approaching them in a positive manner. They then return to play and exploration in the room quickly. They received care at home which can be summarised as being consistently sensitive. In comparison with average levels across all groups, mothers of B type infants were observed at home being more emotionally expressive and provided less contact of an unpleasant nature; at home these infants were less angry and they cried less.

Avoidant infants respond to their carer on reunion in the Strange Situation by not seeking contact or avoiding their carer's gaze or avoiding physical contact with her. These children return quickly to play and exploration but do so with less concentration than secure children. Whilst playing they stay close to and keep an eye on their carer. It may seem that they are not distressed or anxious in the Strange Situation. However, research employing telemetered autonomic data and salivary hormone assays has demonstrate that, despite their relative lack of crying, avoidant infants are at least as

stressed by the procedure as secure and resistant infants (19, page 193). In comparison with average levels across all groups, mothers of A type infants were observed at home being consistently less sensitive to infant signals and less skilled in holding the baby during routine care and interaction. However, in the reunion episodes these infants showed the least anger and crying.

Ambivalent infants respond to their carers on reunion in the Strange Situation by: not being comforted and being overly passive or showing anger towards their carers. These infants do not return quickly to exploration and play. They received care at home which can be summarised as being less sensitive and particularly inconsistent. In comparison with average levels across all groups, C type carers were observed at home being more emotionally expressive; they provided physical contact which was unpleasant at a level intermediate between A and B carers and left infants crying for longer durations; at home these infants were more angry, and they cried more.

Disorganised infants responses in the Strange Situation do not form a clear pattern. Displayed behaviours may be categorised as: contradictory, lacking direction, anomalous, dazed, apprehensive, or disoriented. These infants experience particularly unpredictable and inadequate home caregiving (19, pages 25- 26).

Strange Situation studies provide correlational results that link patterns observed in early carer and infant behaviour with patterns observed in later infant behaviour. Several lines of evidence support an interpretation of these patterns in terms of the infant adapting to their pattern of caregiving they receive, rather than due to innate carer and infant temperaments (28; 19, pages 53-80). Thus for simulation design, infant attachment behaviour at home can be viewed as training data and Strange Situation behaviours as test data. Just as attachment models need to accommodate both the normative findings across episodes as well as the individual differences, they should also take into account that attachment patterns are both sensitive to recent experience (1, pages 217-219) and yet tend to be quite stable across time (19, page 243). If repeated within 10-14 days, infants seem to recognize the test situation, show more distress, more proximity and contact seeking, and the A,B,C patterns are less distinct. And yet the classifications can be quite stable across longer intervals in infancy. The reliability and stability of individual differences in infant-mother attachment predict attachment related measures even in early adulthood (30). An individual's patterns of secure base behaviour therefore remain relatively stable across numerous developmental stages, a noteworthy finding considering the radical changes in an individual's cognitive machinery that occur across these transitions. Perhaps what is common for securely attached individuals across developmental stages is their expectations about their attachment figure's availability and responsiveness in acting as secure-bases. Though the form of these secure-base representations may vary from implicit representations in infancy to secure base scripts in adults (31).

### 3 The Cognitive Component of Bowlby's Attachment Theory: AI and other influences

#### 3.1 Bowlby's conceptual journey

Bowlby published scholarly articles on attachment phenomena between the 1940s and 1980s. During this period the theoretical framework for Attachment Theory was refined and came to incorporate more sophisticated information processing concepts:

*"The hypothesis proposed represents a development of that advanced by me in 1958. The principal change is due to better understanding of control theory and to recognition of the*

*very sophisticated forms that behavioural systems controlling instinctive behaviour may take. In the present version of the hypothesis it is postulated that, at some stage in the development of the behavioural system responsible for attachment, proximity to mother becomes a set-goal. In the earlier version of the theory five patterns of behaviour - sucking, clinging, following, crying, and smiling - were described as contributing to attachment. In the new version these same five patterns are still held to be of great importance, but it is postulated that between the ages of about nine and eighteen months they usually become incorporated into far more sophisticated goal-corrected systems. These systems are so organised and activated that a child tends to be maintained in proximity to his mother."*

*"The earlier version of the theory was described as a theory of component instinctual responses. The new version can be described as a control theory of attachment behaviour" (6, page 180.)*

Bowlby's new version of Attachment Theory shows the continued importance of secure base behaviour with an increasing role for mental representation. However, the new theory still includes a strong ethological influence. For example, Bowlby still presents the attachment system as an instinct to form bonds and as a system that is activated by species specific patterns of care. However, this new version also emphasises three additional components. As a control theory there is a greater focus on the attachment system as directed towards outcomes as set-goals to be achieved from a flexible behavioural repertoire (rather than a system that simply involves triggering pre-set responses). This control system is also described as possessing a sophisticated variety of algorithms, representations, and architectural detail, such as hierarchical structures. Also, the control system as proposed by Bowlby is not just preformed and waiting to be triggered or maturing without experience, but is rather constructed - through interaction between infant learning abilities and information available in the structure of the caregiving environment. With such information processing concepts as these at its core, attachment theory was well positioned to exploit advances in AI. (6; 26).

### **3.2 The attachment control system carries out species specific functions and is inherently motivated.**

Although the control systems formulation was a major departure from Bowlby's early instinct theory, he retained his commitment to behavioral biology. The theoretical inheritance shouldn't be underplayed, as Hinde notes:

*"The concept of a behavioural system is, in fact related to one meaning of the term instinct. [...] It has been used in a rather special sense by ethologists to refer to systems postulated as controlling a group of behaviour patterns that together serve to achieve a given biological end" (Hinde 1983, page 57).*

In animal ethology, behaviour systems are theorised as controlling behaviours such as mating, fighting and feeding. Each behaviour system carries out a species specific function, and has been selected for this function in the evolutionary past. Bowlby suggested that infants possess a somewhat similar species-specific behaviour system that lead to predictable outcomes which are likely to contribute to reproductive fitness. The behaviour systems that Bowlby linked to attachment behaviour in human infants are the attachment, fear, sociability and exploration systems (6). For Bowlby, behaviours resulting from

the attachment behaviour system and the fear system have the predictable outcome of maintaining access and proximity to its primary carer - its secure base and haven of safety. The exploratory behaviour system activates behaviours that result in learning and the sociable system results in social interaction.

A second common thread in the evolution of Attachment Theory is that the behaviour systems most closely related to attachment are inherently motivated. Infants will work to experience exploration, socialisation and security because these outcomes can be considered primary drives. They are not activated as the by-product of any more fundamental process. This means that infants' interest in exploration is a primary motive, not derived from feeding or contact comfort (6). Actions such as running away, freezing and using a carer as a secure base are all behaviours that humans and other animals seem to instinctively 'know' how and when to do when faced with particular dangers. However, this does not mean that these behaviour patterns are not themselves constructed from more basic behavioural components, as we shall see in section 3.5.

### **3.3 The attachment control system uses a flexible repertoire of behaviours in pursuit of set-goals.**

According to Bowlby, what defines the attachment control system is not a set behaviour repertoire but the outcomes that predictably follow from these behaviors (6). Similar behaviours may be produced by different behaviour systems. For example, behaviours such as locomotion can serve more than one system, such as the attachment and exploratory system. Also any given behaviour system may produce a wide range of differing behaviours. In the attachment system, if the infant's goal is to increase its proximity to a carer the infant may cry (which predictably brings the carer closer), or crawl towards the carer themselves. This is an example of behaviours within systems being interchangeable with other functionally equivalent behaviours.

In this formulation, the behaviors used to gain and maintain proximity range from subtle signals, such as gazing towards a carer, to overt signals, such as calling a carer, and active behaviours, such as locomotion. Exploratory behaviours range from locomotion to object manipulation and have the predictable outcome of improving the infant's ability to manipulate the external world.

### **3.4 The attachment control system involves a hierarchy of forms of information processing**

In describing Behaviour Systems, Bowlby (6) invoked a hierarchy of information processing tools which include:

- Ethological concepts and mechanisms, such as Behaviour Systems, Reflex Actions and Fixed Action Patterns which can interact in complex ways by chaining and alternation;
- Concepts from Control Systems theory such as feedback and goal directed mechanisms;
- Concepts from AI and Cognitive Science such as Internal Working Models (IWM's) and hierarchical organisation and control of behaviour using complex representational forms such as hierarchical plans, and natural language.

Reflexes are behaviours with a highly stereotyped form. Once activated by a stimulus at a specific threshold they are ballistically carried to completion. Fixed action patterns are similar to reflexes because they are stereotyped but differ from reflexes because they are open to learning. The thresholds for activation and termination adapt according to the organism's state and past experiences. Fixed action

patterns are also less ballistic, with for example, proprioceptive feedback during execution. Examples include grasping, crying and smiling. Reflexes predominate in the first few months after birth and fixed action patterns predominate from three months until the middle and end of the first year. The reflexes and fixed action patterns that infants perform may seem a very simple form of control but are highly effective in eliciting adult actions that benefit the infant. Different reflexes and fixed action patterns are coordinated together. These complex patterns produced by fixed action patterns can be mistaken for behaviours directed by goal corrected mechanisms because of the sensitive matching of response to stimuli.

The final stage in attachment development, emergence of what Bowlby called a goal-corrected partnership, commences from the middle to the end of the first year. When a mechanism is goal-corrected it is updated or retaken according to feedback on how well the goal has been satisfied. When several goal corrected steps are chained together, and each step must be completed before the next step is taken what has been formed is a plan. Bowlby considered simple plans, and more complex plan hierarchies. He discussed the substructures from which more complex goal corrected plans might be constructed. He also emphasised that, as with orders in a military operation, the attachment control systems involve producing high level plans that are then translated into lower level plans. However, within some limits the lower level subplans can be varied, leading to the flexibility in behaviour repertoire that has already been mentioned above in section 3.3.

Bowlby adapted the working models concept to a more restricted attachment specific concept of Internal Working Models (IWM's), which represent models of self and other in attachment relationships<sup>3</sup>. In this usage, IWM's are held to capture the relation-structure of attachment phenomena, not every aspect of reality but enough that the child can formulate and choose among. These include spatio-temporal causal relations among the events, actions, objects, goals and concepts represented. IWMs represent attachment related world knowledge and expectations about its caregiver's availability and responsiveness. These expectations are derived from the carer's past performance. IWM's of self and attachment figure develop in a complementary manner. An important challenge for current attachment research and theory is to specify in greater detail the cognitive architecture and content of IWMs at different ages. For example if the carer is responsive the self is valued.

Bowlby emphasised the importance of updating IWM's of self and environment:

*"To be useful both working models must be kept up-to-date. As a rule this requires only a continuous feeding in of small modifications, usually a process so gradual that it is hardly noticeable. Occasionally, however, some major changes in environment or organism occur: we get married, have a baby, or receive promotion at work; or, less happily, someone close to us departs or dies, a limb is lost, or sight fails. At those times radical changes of models are called for. Clinical evidence suggests that the necessary revisions of models are not always easy to achieve. Usually they are completed but only slowly, often they are done imperfectly, and sometimes not at all."* (6, page 82)

Reflecting on how IWMs are updated, Bowlby compared the

IWM's of his Attachment Theory with the 'internal worlds' of traditional psychoanalytic theory. He pointed out that that both perspectives assign mental representation a central role in the origins of psychopathology. He notes that the pathological sequelae of separation and bereavement can be understood in terms of models that are partly out of date or full of inconsistencies and confusions (6, page 82).

According to Bowlby, natural language is the ultimate and most sophisticated way in which an individual can represent himself within their social environment. This form of representation has the benefit that *"instead of each one of us having to build his environmental and organismic models entirely for himself, he can draw on models built by others"* (6, - page 82)<sup>4</sup>. A benefit of the non-communicative aspect of language is that the possession of language allows more flexible and imaginative plans and subplans to be created.

### 3.5 The attachment control system is constructed and reconstructed in a developmental process throughout the lifespan.

Bowlby's view of instinctual behaviour changed significantly over the 30 years he devoted to attachment theory. In his final presentation of Attachment Theory this concept was entirely congruent with a developmental theory of control system construction. According to Bowlby, the "instinctual" behaviors evolution provides are not limited to simple reflexes. They can be complex, goal corrected, and socially relevant. As an example, Bowlby (6, page 44) points to the falcon's ability to capture small birds in flight as a quintessential example of an instinct: the falcon's stoop is species specific and appears largely without opportunity of learning. However, Bowlby, emphasises that even if we don't understand how control systems of this kind come to develop, we shouldn't doubt that there is a developmental explanation that relies upon the Falcon interacting within the ordinary expectable environment for its species.

Bowlby suggests that there are multifarious processes whereby early appearing fragments of instinctive behaviour are integrated into later appearing complete sequences with their normal mature functional consequences. These integrative processes include restrictions appearing in the activating and terminating conditions of the behaviours, and behaviours becoming units in one or more chains. Bowlby also describes a third process:

*"Yet another sort of process [of functional integration] is one that integrates a piece of behaviour within a causal hierarchy. This can occur following a change in the causal relation between a pattern of behaviour and the internal state of the animal."*

*It might confidently be supposed that feeding behaviour would be most readily elicited when an animal is hungry, and the hungrier it is, it might be thought, the more readily would the behaviour be elicited. This is by no means so, however, at least in the very young. For example, when a fledgling great tit starts to peck it is most likely to do so when it is **not** hungry: when it is hungry it begs food from its parents. Similarly, experiment suggests that sucking behaviour in puppies is at first independent of both hunger and food intake. Later in development pecking and sucking become elicited most readily in conditions of hunger and, by those means, are brought, together with other behaviour contributing to food-intake, within a system organised in terms of causal hierarchy."* (6, pages 159-160.)

<sup>3</sup> Kenneth Craik originally introduced the term 'working model'. However, as Bretherton (10) notes Craik's concept of Working Models came to Attachment Theory indirectly from the through the writings of the biologist J. Young (1964).

<sup>4</sup> compare with Dennett's description of Gregorian Minds (17, page 99)

Clearly this process of causal hierarchy formation in great tit chicks doesn't result in the sophisticated kinds of representational redescription characteristic of human social and cognitive development. However, Bowlby goes on to set out a framework for how the kinds of processes that integrate behaviours in chicks and puppies can have far reaching effects in humans:

*"Because of a human's immense capacity to learn and to develop complex behavioural systems, it is usual for his instinctive behaviour to become incorporated into flexible behavioural sequences that vary from individual to individual. Thus once a human has had experience of reaching a consummatory situation the behaviour that leads to it is likely to become reorganised in terms of a set-goal and a plan hierarchy"* (pages 160, 6)

These developmental processes described by Bowlby involve intimate interaction between lower level processes, such as simple reflexive responses, and emerging higher level structures and mechanisms. New resources come online gradually. Integrating elements into a system depends on (i) biases in infant learning abilities and (ii) information/structure in the expectable caregiving environment. Absent the knowledge about the latter social/caregiving affordances, and the former biases in learning abilities are of no use. Adult cognitive abilities are radically different from an infant's, but humans do not undergo some kind of process of metamorphosis like cognitive-caterpillars changing to cognitive-butterflies, where a brand new architecture emerges suddenly. Rather, new resources are broken in slowly, and take shape in an environment where the ongoing behaviour patterns have been set by the older earlier maturing resources.

#### 4 Updating the Cognitive Component: designing and implementing simulations

Despite Bowlby making a valuable contribution towards explaining attachment phenomena in information processing terms, his ideas are not set out as design specifications merely awaiting implementation. He may have referred to ideas in Control Systems Theory and AI as much to break the old Freudian paradigm than to specify in great detail a way forward to more involved modelling and simulation. However, even if this were the case, Bowlby's information processing framework can still inform simulation design, but it must be interpreted, and in some sense extrapolated.

A number of different attachment control architectures have been implemented as simulations. These include simulations of: approach and exploratory behaviour in naturalistic environments (3; 26, pages 51 - 78); infant agents adapting to the secure or insecure patterns of caregiving agents (26, pages 79 - 102); and infant and carer agents reproducing the different behaviour patterns observed in the Strange Situation Experiment (26, pages 103 - 152).

Most of the examples above involve autonomous agent architectures which possess a generic family resemblance. Though they possess multiple components, they can be considered homogenous architectures (in contrast with hybrid architectures) because each component is a single generic type of reactive subsystem (the components being similar to those incorporated in Behaviour Based architectures in robot or autonomous agent simulations (32; 26, pages 62 - 69)). Each of the different component 'behaviours' correspond to one of the behaviour systems that Bowlby invoked as involved in secure-base behaviour (exploration, sociable, attachment-security and wari-

ness). In these homogenous architectures, the highest activated behaviour can control the actions that are carried out by the agent. However, the decision processes is not strictly winner take all because two goal activators with high activation may mutually inhibit each other, leaving a less highly activated goal to direct behaviour, and give the impression of displacement activity. Key attachment phenomena in real infants can be explained by interaction among the same type of information processing components incorporated in these agent based simulations. For example, A type infants have experienced aversive responses at home, particularly in close proximity. The phenomenon of A type infants avoiding their carers in reunion episodes of the Strange Situation can therefore be explained by the postulation of reactive avoidance behaviour systems in these infants becoming more activated than other components in the emotionally aroused environment of reunion episodes.

Interesting response patterns can emerge in simulations with infant agents possessing relatively simple homogenous reactive architectures. For example, computational experiments demonstrate novel epigenetic trajectories for the development of Secure and Insecure attachment behaviour patterns (26; 27). A key parameter in these simulations is the infant agent's 'confidence' in its carer agent's responsiveness. This parameter is represented in the infant agent's architecture as a 'safe-range' separation distance between infant and carer agent that the infant will tolerate without any rise in activation for the attachment anxiety goal activator. When separation goes beyond the infant agent's 'safe range' the attachment anxiety goal activator starts to become activated, and therefore competes with other goal activators for control of setting the next action. The attachment goal activator will become deactivated by the proximity to the carer agent increasing, or by the infant gaining the carer agent's visual attention. Simulations have demonstrated that for certain infant agent architectures with certain initial parameters, the infant and carer agents can interact in a positive feedback process that results in moderate levels of infant confidence in the carer agents responsiveness becoming unstable over lengthy periods of interaction. This instability results in populations of identical infant-carer agent dyads with initially moderate settings for confidence and responsiveness bifurcating into Secure and Insecure subgroups. This is interesting because clustering of attachment groups is found in empirical results from Strange Situation studies (1; 27; 26, pages 79 - 102). Bifurcation occurs because the effect of very small random variations in movements and timings on the otherwise identical infant-carer agent interactions are greatly amplified by the positive feedback. These initially identical agents end up as two groups with no intermediate cases and therefore could be categorised as separate taxons (24; 26, pages 97 - 99).

A contrasting example of a more complex attachment control system that has been implemented is the hybrid architecture with reactive and deliberative subsystems shown in figure 1. This hybrid architecture possesses reactive behaviour based components (similar to those described above in the homogenous architectures above), that represent information implicitly. However, in this architecture these reactive subsystem interact with a higher level deliberative reasoning subsystem that represents goals and the state of the world explicitly. This means that this architecture can construct explicit plans, which are composed of valid operations from the current the state of the world towards desired goal states. Figure 1 shows the higher level deliberative subsystem can interact with the lower system by inhibiting a reactive prepotent response (denoted by the dashed line), when this response may result in an undesirable outcome. Relatively sophisticated processes of reasoning and inhibition such as this may be easily overwhelmed by lower level reactive processes for one year

old infants in the emotionally aroused context of the Strange Situation. However, in older children and adults such reasoning about social and emotional goals may interact with reactive processes and both influence attachment behaviour patterns. However, this particular example of a hybrid architecture is too simple and inflexible to capture the richness of attachment responses in the older age groups.

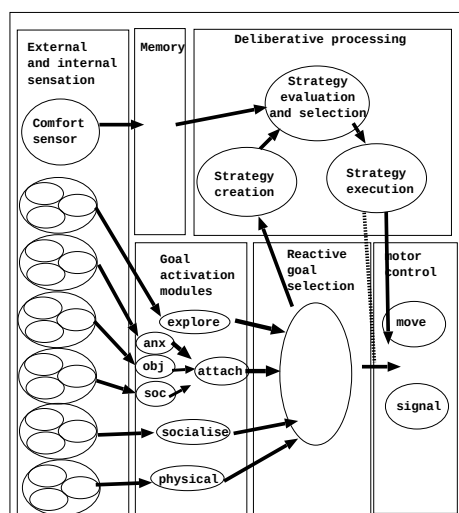


**Figure 1.** A Hybrid Design for the Attachment Control System. Deliberative mechanisms provide a secondary route to action, activated as a result of interrupts to reactive selection and arbitration and imposing different actions. The reactive goals are: exploration (explore); anxiety (anx); object wariness (obj); social wariness (soc); socialisation (socialise); and physical need (physical). The reactive goals of anxiety, object wariness, and social wariness are all combined into a single attachment goal (attach) before they are considered for selection. The comfort sensor measures contact pleasure and represents this information symbolically. Avoidance can occur when the deliberative subsystem inhibits approach because it has reasoned about the undesirability of close interaction.

Reactive architectures that lack the kinds of deliberative resources shown in figure 1 can simulate a wide range of naturalistic and laboratory attachment phenomena observed in infants. However, for older children and adults, deliberative subsystems more sophisticated than that presented in figure 1 would be expected to play some part in producing attachment behaviours. This raises a number of questions. What kinds of subsystems influence attachment behaviour at different ages? How will different subsystems interact? How will the empirical finding of continuity in attachment patterns be supported as higher level subsystems come ‘on-line’? Lastly, how might new subsystems be constructed?

## 5 Future steps - Applying Bayes to Bowlby?

Dayan and colleagues (12; 15; 16; 22) have developed a theoretical framework that may help provide traction in answering some of these questions. This framework includes four action controllers, labelled: Pavlovian; Habitual; Model-based; and Episodic (13). These four controller possess complementary strengths and weaknesses and their operation is integrated together. In the context of attachment, a coherent behaviour pattern might be driven by more than one controller. For example, the initially observed behaviours in a Strange Situation episode might produced by the Pavlovian controller (whose

responses are likely to be fastest). Other controllers may exert their influence later in the episode, but still be recognised as within the same attachment pattern response.

Recent evidence suggests how these controllers may be instantiated in specific brain areas (12; 13; 16; 22). In addition, the responses of the Habitual, Model-based and Episodic controllers are considered to be mediated by a decision principle based upon uncertainty (12; 13; 22). Each of these three controllers incorporates a measure of its own uncertainty in its likely performance. The least uncertain controller is then chosen to direct the next action. This Bayesian decision mechanism may account for the Disorganised (type D) pattern of infant response in the Strange Situation. This pattern of response arises when infants receive unpredictable maternal care. One can imagine, for example, a Bayesian arbitration system producing “disorganized” behaviour in the face of unpredictable maternal care. In this case no controller would reach the level of certainty needed to direct action with coherence and consistency.

### 5.1 The Pavlovian controller

This controller performs routine and reflex behaviours such as fleeing from danger (15). Unlike the other three controllers the Pavlovian controller does not rely upon the individual learning from experience to adapt how it responds. Rather this controller possesses a set of responses that are tied to triggers that have been set in an evolutionary process. The reflexes that Bowlby described as being rudimentary parts of the attachment control system can be considered as part of the Pavlovian control system. Allowing this system to control behaviour affords speedy response and requires minimal cognitive load. However, its major disadvantage is that it is inflexible, and its responses are only adaptive if you current environment matches that in which its settings were evolved. The avoidant behaviour typical of A-type infants in the Strange Situation, and of adolescents and adults in conflict with romantic partners (19, page 179), may arise from triggering Pavlovian aversive responses (16).

### 5.2 The Habitual controller

The Habitual controller, in common with the Pavlovian controller, exerts its influence on our behaviour outside of our conscious awareness. As with the Pavlovian Controller, it is a relatively fast system, and also involves minimal cognitive load. However, unlike the Pavlovian controller, it can learn from the results of its previous attempts to gain a reward or avoid a punishment in particular environmental states, and might therefore be compared to the fixed action patterns described by Bowlby. A drawback of the Habitual controllers learning mechanism is that its initial attempts to accomplish a task are often far removed from the optimal performance that it can ultimately gain after extensive experience. Where Model-based control (described in the next section) may include all available information to decide which action to take, habitual control may use just a single simple metric that represents the overall utility of a taking a particular action for a particular state of the environment. This cached value can be seen as comparable to what would be gained if the total anticipated value of all future actions were collapsed to a single value (16). The Habitual Controller has a notable drawback - that the cached values are not open to reflection and cannot be changed in a ‘one-shot’ manner when the environment changes or when new information comes to light. Since attachment behaviours only change gradually, and people often struggle to reflect on their attachment

related actions (19, 43-46), it may be that much of our attachment behaviour is under Habitual control.

According to the Bayesian framework, actions directed by the Pavlovian control system can help shape or interfere with instrumental responses arising from the habitual control system (15). It may be that Pavlovian Instrumental Transfer (PIT) from biases in frequently activated reflex responses to the Habitual Controllers causes the observed continuity in attachment behaviour patterns. Processes that bring controllers into indirect coordination without direct communication may occur throughout the life-span and bring Pavlovian and Habitual controllers into congruent attachment behaviour patterns in a process outside of an individual's awareness.

### 5.3 The Model-based controller

This controller (also described as the goal-directed controller) involves conscious reasoning in pursuit of explicit goals (14; 18). It can construct plans, representing all its alternative choices and their consequences as models of the self and environment. As it models the environment, this controller can use all available information about the task and environment in its calculations, and when something changes in the environment, it can immediately incorporate this new information in a statistically efficient manner in its model. It therefore has immediate and optimal access to the results of experience. However, this very flexibility leads to its major disadvantages. Considering all available data can be overly time consuming and prohibitive in computational load for all but the simplest problems.

It is likely that Model-based control does not have a central role in controlling infant attachment behaviour of infants, but explicit reasoning about what actions to take within adolescent and adult relationships, including caregiving, seems a reasonable proposal. This might be construed as implementing Bowlby's important aspects of working models concept. It might also prove useful in modeling the goal corrected partnership. However, the reward focused Habitual Controller might also be recognised as a rather minimal form of IWM as it predicts optimal actions from experience.

It was noted above that the observed continuity of attachment behaviour patterns through the life-span might arise when Pavlovian responses shape Habitual behaviour. Might an analogous process occur between the unconscious processing by Pavlovian and Habitual controllers, and attachment behaviour that resulted from the operation of the Model-based controller? The issue here is how does implicit knowledge become explicitly represented. One possibility is that individuals observe and recognise their own unconsciously produced behaviour patterns. They may then elaborate upon these and use them in building their own personal narratives that are congruent with the original behaviour patterns. Thus implicitly representations directing Pavlovian and Habitual patterns of behaviour may not directly turn into explicit representations that are held within the Model-based system. Rather, the unconscious controllers may drive behaviour patterns that are then observed in their interaction with the environment, and it is an individual's interaction with their environment that is observed by their conscious mind, in form of the Model-based controller.

If response patterns arising in the Pavlovian and Habitual controllers may influence response patterns in the Model-based controller, the reverse may also occur. The Model-based system may provide a kind of 'will power' for individuals to overcome their own unconsciously produced responses that they recognise and wish to minimise (15). For example, in adult romantic relationships some individuals (perhaps after advice) may attempt to limit their habitual

avoidance of intimacy. Whereas others individuals, who often become emotionally enmeshed, may recognise the benefit of deferring intimacy (19, pages 178-180).

### 5.4 The Episodic controller

The Episodic controller fits a gap between the Model-based and Habitual controllers (22). This might be when reasoning is too slow, or requires too many cognitive resources, for the Model-based controller to operate effectively, and in addition, the situation that has not been experienced enough times for the Habitual response to be optimised. This controller simply chooses an action that gained a favourable response in a similar previous situation, even if this situation were experienced rarely. Recalling the sensitivity of infant behaviour to context when the Strange Situation is repeated within 10-14 days, this controller might therefore be a good candidate explanation for infant behaviour when re-tested in the Strange Situation after only a few weeks. In these cases the infant is in an unusual situation that has only been experienced in this precise manner once before. Since in the original Strange Situation all infants show some level of separation protest, and the carers always returned, it may be an Episodic controller that activates increased separation protest when this procedure is repeated.

### 5.5 Self-constructing control systems

In addition to explaining how behavioural patterns are transferred between subsystems (which already exist), one of the key future directions for research is how subsystems in the attachment control system actually become constructed. In human subjects it may be that a deliberative system similar to the Model-based controller starts to increasingly direct attachment behaviours from late infancy through to adulthood. As it does so, the form of the IWM's and other representations that it utilises, such as natural language, will become more sophisticated. These higher level representations may then provide a richer substrate for the acquisition of new habitual skills in influencing an individual's attachment relationships (18).

Hierarchical structure in an Habitual controller may thus develop following the employment of hierarchically structured actions by the Model-based controller. However, how does hierarchical structure develop within the Model-based controller? As described in section 3.5, causal hierarchies related to the goal of feeding can develop from actions such as pecking in chicks or sucking in puppies. In these examples, the pecking or sucking actions are initially unrelated to the pursuit of food. From the broad initial range of behaviours (that might be determined in the genome), only a narrower range is selected to be 'set' as feeding behaviours that predictably result in gaining food. So later in development only this subset of the original behaviours is triggered by food stimuli. In altricial species it may be common that particular sets of genome determined behaviours are only recognised post-natally as possessing causal potential. Delaying acquisition of associations between action and goal may be repeated numerous times in a cascaded developmental process (11). This process can then lead to hierarchical action control structures forming as actions and goals are represented at greater levels of granularity (11).

## 6 Conclusion

This paper has described a range of attachment behavioural phenomena and presented John Bowlby's (6; 7; 8) explanatory framework for these phenomena. It has then surveyed recent autonomous agent

simulations which have built upon information processing aspects of Bowlby's framework. These simulations have highlighted limitations in the current understanding of how subsystems within the overall attachment control system may interrelate and become constructed through development. Two AI frameworks, one Bayesian (14) and the other concerned with cascaded hierarchical development (11) have been presented that might inspire developmental explanations of attachment. Key benefits of the Bayesian framework are threefold. In this approach: attachment behavioural phenomena can be mapped onto the performance of constituent subsystems that have been implemented in simulations of other behavioural phenomena; the subsystem interactions in simulations of the attachment control system can be decided in a principled 'Bayesian' manner; and lower level reactive components can 'shape' the nature of patterns of attachment response in higher level deliberative components. Simulating attachment development as an example of a cascaded developmental process may shed light on how the hierarchical structure of the Attachment control system is constructed. Contemporary measurement tools allow attachment response patterns observed in infancy to be compared with those exhibited in adulthood (31). Computational modelling with inspiration from AI may provide new explanations for observed continuities in patterns of attachment response across the life-span.

## REFERENCES

- [1] M. Ainsworth, M. Blehar, E. Waters, and S. Wall, *Patterns of Attachment: a psychological study of the strange situation*, Erlbaum, Hillsdale, NJ, 1978.
- [2] J.W. Anderson, 'Attachment behaviour out of doors', in *Ethological Studies of Child Behaviour*, ed., N. Blurton-Jones, 199–215, Cambridge University Press, Cambridge, UK, (1972).
- [3] N. Bischof, 'A systems approach toward the functional connections of attachment and fear', *Child Development*, **48**(4), 1167–1183, (1977).
- [4] J. Bowlby, 'Forty-four juvenile thieves: Their character and home life.', *International Journal of Psychoanalysis*, **25**, 1–57, (1944).
- [5] J. Bowlby, 'Grief and mourning in infancy and early childhood', *The psychoanalytic study of the child*, **XV**, 9–52, (1960).
- [6] J. Bowlby, *Attachment and loss: volume 1 attachment*, Basic books, New York, 1969. (Second edition 1982).
- [7] J. Bowlby, *Attachment and loss: volume 2, Separation: Anxiety and Anger*, Basic books, New York, 1973.
- [8] J. Bowlby, *Attachment and loss: volume 3 loss, sadness and depression*, Basic books, New York, 1980.
- [9] J. Bowlby and J. Robertson, 'A two year old goes to hospital', *Proceedings of the Royal Society of Medicine*, **46**, 425–427, (1952).
- [10] I. Bretherton, 'The origins of attachment theory: John bowlby and mary ainsworth', *Developmental Psychology*, **28**(5), 759–775, (1992).
- [11] J.M. Chappell and A. Sloman, 'Natural and artificial meta-configural altricial information-processing systems', *International Journal of Unconventional Computing*, **3**(3), 211–239, (2007).
- [12] N.D. Daw, Y. Niv, and P. Dayan, 'Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control', *Nature Neuroscience*, **8**, 1704–1711, (2005).
- [13] P. Dayan, 'The role of value systems in decision making', in 'Better than Conscious? Decision Making, the Human Mind, and Implications for Institutions', eds. C. Engel and W. Singer, 51–70, MIT Press, Frankfurt, (2008).
- [14] P. Dayan, 'Goal-directed control and its anitpodes', *Neural Networks*, **22**, 213–219, (2009).
- [15] P. Dayan, Y. Niv, B. Seymour, and N.D. Daw, 'The misbehaviour of value and the discipline of the will', *Neural Networks*, **19**, 1153–1160, (2006).
- [16] P. Dayan and B. Seymour, 'Values and actions in aversion.', in 'Neuroeconomics: Decision making and the brain', eds. P. Glimcher and C. Camerer and R. Poldrack and E. Fehr, 175–191, Academic Press, New York, (2008).
- [17] D.C. Dennett, *Kinds of minds: towards an understanding of consciousness*, Weidenfeld and Nicholson, London, 1996.
- [18] K. Douglas, 'The other you', *New Scientist*, **196**(2632), 42–46, (2007).
- [19] S. Goldberg, *Attachment and Development*, Arnold, London, 2000.
- [20] 'H. Harlow, 'The Nature of Love', *American Psychologist*, **13**, 573–685, (1958).
- [21] J. Holmes, *John Bowlby and Attachment Theory*, Routledge, 1993. (revised edition).
- [22] M. Lengyel and P. Dayan, 'Hippocampal contributions to control: The third way', in *21st Annual Conference on Neural Information Processing Systems*, pp. 1–8, (2007).
- [23] K. Lorenz, *King Solomons Ring*, Methuen, London, 1952.
- [24] P. Meehl, 'Clarifications about taxometric method', *Applied and Preventative Psychology*, **8**, 165–174, (1999).
- [25] D. Petters, 'Simulating infant-carer relationship dynamics', in *Proc AAAI Spring Symposium 2004: Architectures for Modeling Emotion - Cross-Disciplinary Foundations*, number SS-04-02 in AAAI Technical reports, pp. 114–122, Menlo Park, CA, (2004).
- [26] D. Petters, *Designing Agents to Understand Infants*, Ph.D. dissertation, School of Computer Science, The University of Birmingham, 2006. (Available online at <http://www.cs.bham.ac.uk/research/cogaff/>).
- [27] D. Petters and E. Waters, 'Epigenetic development of attachment styles in autonomous agents', in *Proceedings of the Eighth International Conference on Epigenetic Robotics. Modeling Cognitive Development in Robotics Systems*, eds., M. Schlesinger, L. Berthouze, and C. Balkenius, 153–154, Lund University Cognitive Studies, 139, (2008).
- [28] L.A. Sroufe, 'Attachment classification from the perspective of infant-caregiver relationships and infant temperament.', *Child Development*, **56**(1), 1–14, (1985).
- [29] E. Waters, K. Kondo-Ikemura, G. Posada, and J. Richters, 'Learning to love: Mechanisms and milestones', in Minnesota Symposium on Child Psychology (Vol. 23: Self Processes and Development), eds. M. Gunner & Alan Sroufe, 217–255, Psychology Press, Florence, KY, (1991).
- [30] E. Waters, S. Merrick, D. Treboux, J. Crowell, and L. Albersheim, 'Attachment stability in infancy and early adulthood: A 20-year longitudinal study', *Child Development*, **71**, 684–689, (2000).
- [31] H. Waters and E. Waters, 'The attachment working models concept: Among other things we build script-like representations of secure base experiences', *Attachment and Human Development*, **8**(3), 185–197, (2006).
- [32] M. Wooldridge, *An introduction to multiagent systems*, Wiley, chichester, 2002.

# Construct to Understand : Learning through Exploration

Zoe Demery<sup>1</sup>, Veronica E. Arriola Rios<sup>2</sup>, Aaron Sloman<sup>2</sup>, Jeremy Wyatt<sup>2</sup> & Jackie Chappell<sup>1</sup>

**Abstract.** Artificial Intelligence (AI) and Animal Cognition (AC) share a common goal: to study learning and causal understanding. However, the perspectives are completely different: while AC studies intelligent systems present in nature, AI tries to build them almost from scratch. It is proposed here that both visions are complementary and should interact more to better achieve their ends. Nonetheless, before efficient collaboration can take place, a greater mutual understanding of each field is required, beginning with clarifications of their respective terminologies and considering the constraints of the research in each field.

## 1 INTRODUCTION

A list of advantages and problems within each field is presented. It is proposed that AI and AC can complement each other and help in solving issues that have proved hard to tackle separately within each discipline. For example, within AI it is known how the 'robot brain' was coded and therefore we know what it does and how it does it. In contrast, a wide variety of behaviours can be observed in animals, but they are hard to analyse since there is little direct information about the structures and processes taking place within their brains. The implementation of concrete structures and processes in a computer model, where the robot's behaviour can be compared to an animal's behaviour, opens a new route to propose models for AC that could not be tested in another way.

Before collaboration can take place, a number of problems must be addressed. For instance, AC must overcome its anthropocentric view and clarifications must be made where similar terms are used in AI and AC but have different meanings. Scientists of both fields must keep an open mind towards new techniques and different, unfamiliar views on actually similar questions asked in their own field (Figure 1).

The poster further presents the study of how an individual learns about the deformability of objects as an example to where AI and AC can complement each other. We propose the Three-stage Theory of Exploration (Forming, Testing and Extending & Refining hypotheses), which can be studied from opposing directions by both AC and AI. While AC presents hypotheses about the behaviour that can be observed in New Zealand red-fronted parakeets (*Cyanoramphus novaezelandiae*), AI proposes models that are expected to produce analogous observations in robots<sup>1, 2</sup>. Both views are summarised in parallel boxes where similarities and differences are highlighted.

<sup>1</sup>School of Biosciences, Univ. of Birmingham, B15 2TT, UK.  
Email: [zxd878@bham.ac.uk](mailto:zxd878@bham.ac.uk).

<sup>2</sup>School of Computer Sciences, Univ. of Birmingham, B15 2TT, UK.  
Email: [v.arriola-rios@cs.bham.ac.uk](mailto:v.arriola-rios@cs.bham.ac.uk).

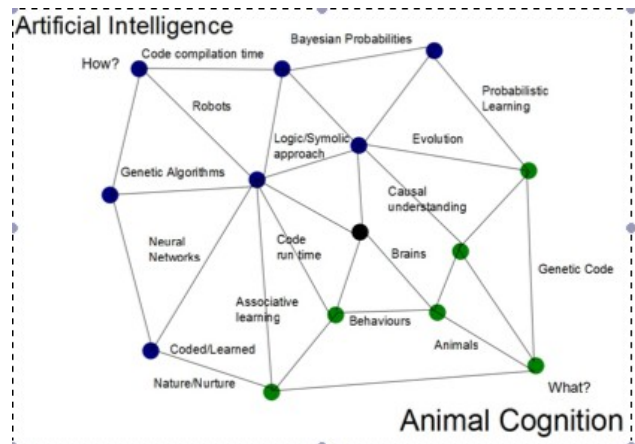


**Figure 1.** A mind map illustrating how different important concepts and terms in Artificial Intelligence (blue) and Animal Cognition (green) can link to provide a fuller answer to a research question posed by both fields (black)

## 2 COMMON PROBLEMS & THEIR INTER-DISCIPLINARY ANSWERS

Firstly, both fields are traditionally quite anthropocentric. It is often thought that there is a dichotomy between the cognition of humans and non-humans animals. However, despite each individual and species having unique mechanisms to cope with their unique ecological niche, it is worth investigating competences developed by individuals other than humans. In modelling cognition a single separated type of model is not enough and collaboration between different mechanisms may be required.

Secondly, quite frequently the same terms in each field refer to very different (although analogous) concepts, or similar concepts are called different names. For instance, in AC 'neural nets' are connections between neurons in the brain. While in AI they are numerical techniques to approximate an unknown function, inspired by biological systems, but they really do not do the same function and are a lot simpler.

Both fields can help each other if researchers are open to learning and understanding new ideas and terminology proposed by the other field that can seem initially very daunting. While traditional views are always the foundation of any field, techniques, approaches and knowledge are constantly changing and evolving.

There is always more than one way to phrase a question and considering different angles of the situation can provide a more complete answer. This is why collaboration is fruitful and productive. AI & AC are complementary sides of the same coin.

For instance, AC often follows the *top-down* approach when trying to answer a research question (what are the requirements imposed by the environment that results in this particular behaviour?), while AI follows the *bottom-up* approach (let's try to build the system from scratch). At some point, they meet in the middle and provide a complete understanding of the interesting behaviour or learning strategy in question.

### 3 AN EXAMPLE PROBLEM: HOW DOES AN INDIVIDUAL LEARN ABOUT THE DEFORMABILITY OF OBJECTS?

We have combined our respective expertise in AC and AI to propose a 'Three-stage Theory of Exploration' to explain how individuals learn about the physical properties of their surrounding environment. Specifically, we apply it to a frequent object affordance or property experienced in everyday life – the rigidity or deformability of objects.

Today AI and AC researchers either seem to be focussed on how learning is by simple associative trial-and-error, or probabilistic and statistical strategies (e.g. Bayes nets) or by more complex and abstract causal understanding. We propose that it is in fact by a combination of these strategies depending both on the conditions of the environmental situation and the competences of the individual (whatever species). Surprisingly little research has looked at how systematic exploration could support learning strategies.

We believe that exploration is not always random but structured, selective and sensitive to particular features and salient categorical stimuli of the environment. It typically follows through three stages of theory formation specific to particular affordances and processes, but can be generalised to novel situations. Firstly, individuals form internal representations of the world/environment by acting on it in a way that suggests probabilistic reasoning and simple trial-and-error learning is being used. Their exploration further reflects genetic predispositions to particularly salient environmental features, such as edges and corners. Moreover, they perhaps represent and segregate the world into general categories from birth. Secondly, the individual tests these internal hypotheses. They progressively use more complex mechanisms like causal reasoning, so explorative actions become less repetitive and they generally behave more flexibly depending on the situation they are presented with in their environment. Tests internal theories and rules of object relations by performing (initially) certain actions on certain objects. Lastly, individuals extend and refine these hypotheses by analysing their theories and reusing knowledge about the world by combining them in similar but new environmental situations. The individuals can also now fill in gaps in knowledge abstractly by causal inferences, so hypotheses are extended throughout life.

We realise that at this stage, this theory is very general and still has a lot to be expanded upon and considered in more detail. However, it is an example of how apparently different forms of thinking from different research fields can actually be brought together and thus provide plenty more scope for collaboration and further more systematic investigation. It is purposefully designed in this way also to provide a more general framework so that it can be applied to a wide range of behaviours, species and environments. However, we do acknowledge that there are a

variety of contributing factors and thus different specific strategies and mechanisms for different contexts. Thus in the next section we briefly explore our specific hypotheses for this context – learning about the deformability of objects in a robotic agent and parakeets.

### 4 PROPOSED METHODS & HYPOTHESES

When AC and AI work in parallel the order and priority of their respective hypotheses often becomes altered, yet the methods and results are complimentary. In the case of deformable objects the methods for both can be compared as follows. From the AC perspective, we plan to consecutively present individual parakeets with novel objects of different shapes and deformabilities. Then we will record in detail their actions from a rigid ethogram specific to this behavioural context, backed up by video analysis.

However, from the AI point of view, we have taken a more bottom-up approach and propose a set of candidate mass spring models for every material the individual (i.e. robot) encounters (Figure 2). Subsequently, we will search for the model that best fits the data that the robot acquires through selective interactions with the material. Then we will detect when no model can offer a 'good' representation and evaluate and characterise the search strategies that find the best models and use these results to improve the searching process. Finally, the system will be tested by consecutively presenting the robot with simple solid and deformable novel objects in a similar task design to the parakeets. The quality of the generated models and the number of steps (time) it took to find them will then be evaluated.

It can hence be observed that while AC focuses on the observation of behaviours, AI focuses on the construction of the infrastructure that would produce those behaviours. The hypothesis in each case reflects the opposing goals of both disciplines and the application of the Three-stage Theory of Exploration.

Firstly, in AC we predict that there will be a higher frequency of the parakeets touching the corners and areas of high curvature over the smooth surfaces. Furthermore, they will explore a novel deformability more than a familiar one. Then between trials, the individual begins to explore the most compressible and the most rigid objects more than the other levels of intermediate deformability, but they progressively focus in more detail on these intermediate levels. Lastly, we hypothesise here that the parakeets will eventually reach a point where they have so much experience with the different objects' compressibility and affordances, that they can combine their understanding of the different objects to solve a new task. However, how this will be observed specifically in the behaviour is still to be decided.

On the other hand in AI, we hypothesise that it is possible to generate a geometrical internal representation on top of which a mass spring physical model can be applied<sup>[1][2]</sup>. The nodes of the mesh will be more likely to be placed at points of high curvature. Secondly, we believe that it is possible to calibrate the parameters of the model by using a random search algorithm. This algorithm samples the space of possible parameters and selects the most suitable ones. Finally, we predict that it is possible to improve learning by providing a clustering algorithm that learns to segment and classify the space of parameters of the physical model, so that the appropriate models for new materials are calibrated faster.

Consequently, from these two sets of hypotheses we can see that in parts they are actually very similar, and where they are not so similar, they can add resolution to each other. For instance, it can be seen that while the first hypothesis of AC looks for evidence of more attention on places with high curvature, the first hypothesis of AI directly establishes that the model will use points of high curvature as pivotal points for the model and expects that the generated model of the behaviour of the deformable object will be good.

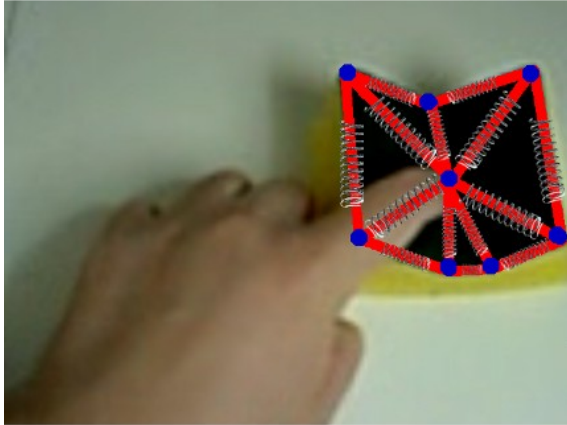


**Figure 2.** In Artificial Intelligence a generic technique from physics (mass-spring models) is used to generate an internal representation of deformable objects

## 5 CONCLUSIONS & FUTURE WORK

For AI, the problem of interest is to learn to predict the behaviour of deformable objects based on experience. In order to achieve that, a learning technique is proposed, where the generation and selection of physical models consists of sets of masses connected by springs on top of a mesh of triangles superimposed onto the object of interest. From an AC point of view, it is hypothesised that this learning technique will produce in the robot a behaviour that is analogous (though perhaps simpler) to what can be observed in the parakeets when encountering novel deformable objects. Although parakeets may not directly use a mass-spring model to learn about the deformability of their environment, they may use a similar frequency-based strategy that can be derived from the mass spring model. New hypotheses and experiments could then be designed to further test the applicability of this model and to hopefully extend it.

## REFERENCES

- [1] D. Morris. Automatic Preparation, Calibration, and Simulation of Deformable Objects. *Computer Methods in Biomechanics and Biomedical Engineering*, 11(3):263-279.
- [2] M. Teschner, B. Heidelberg, M. Muller, and M. Gross. A Versatile and Robust Model for Geometrically Complex Deformable Solids. In *Computer Graphics International (CGI 2004)*. pp. 312-319 (2004).

# The PoulBot: a mobile robot for ethological studies on domestic chickens

Alexey Gribovskiy and Francesco Mondada<sup>1</sup> and José Halloy and Jean-Louis Deneubourg<sup>2</sup>

**Abstract.** This paper presents the “Mixed Societies of Robots and Vertebrates” project that is carried out by the Mobile Robotics group at EPFL, Switzerland in collaboration with the Unité d’Ecologie Sociale at ULB, Belgium. The idea of the project is to study behavior in young chicks of the domestic chicken (*Gallus gallus domesticus*) by using mobile robots able to interact with animals and accepted as members of an animal group. In contrast to other studies where relatively simple robotic devices are used to test specific biological hypotheses, we aim to design a multipurpose robot that will allow ethologists to study more sophisticated phenomena by providing a wide range of sensors and sufficient computational power. In this paper we present a PoulBot robot and give an example of the ethological experiment, where it is used to study the social aggregation in chick groups.

## 1 INTRODUCTION

One of the long-standing interests in behavioral studies is to understand relationships between stimulus and response. To study this mechanism researchers often use specially designed mock-ups whose behavior can be controlled in details to monitor a response of the animals under test. Being originally quite simple, nowadays such artificial models become sophisticated robotic devices thanks to availability of low-cost miniaturized computer chips, motors and sensors. This allows to test hypotheses on animal behavior that were to tricky to work with before [5]. Recently robots were used to test ideas about the mate selection in bowerbirds [7] and in tungara frogs [9], to study the male territorial instinct in dart-poison frogs [6], the nest mate recognition in brush turkeys [1], the predator avoidance by ground squirrels [8] and the rats cognitive ability [4]. In most of these studies robots were built to test only one or very few specific behaviors and thus endowed with a limited number of sensors and computational power; the majority of them had to be remotely controlled during the whole experiment.

The goal of our work is to develop a more advanced mobile robot, equipped with a wide range of sensors (cameras, microphones, proximity sensors, etc.) and a sufficient computational power, which will allow ethologists to study more sophisticated phenomena. This robot will be used to address such scientific questions as link between individual and collective behavior, sound communication, role of personality in collective decision making and pattern recognition in the domestic chicken (*Gallus gallus domesticus*). We also model the collective movement behavior. The individual based models that are usually used to model birds flocks are not generally applicable for

small groups, since they consider all the animals to be identical and groups being large. We, on the contrary, aim to research how individual features and leadership influence the collective movement.

Methodologically our project continues the European project Leurre, where mobile microrobots were designed to study and control the social aggregation in cockroaches [2]. Another example of a project dealing with domesticated birds-robot interaction is the Robot Sheepdog Project where a mobile robot was designed to shepherd a flock of ducks and to lead them to a specified position by scaring them [10].

In this paper we present the current progress in the project: the PoulBot robot design and control system, and experiments on the social aggregation in chick groups that we are currently conducting.

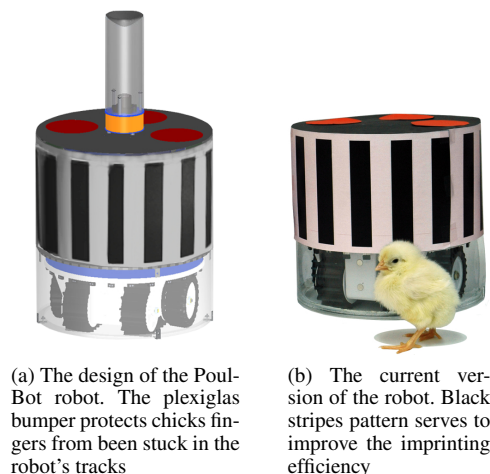


**Figure 1:** The PoulBot design and its current prototype

## 2 THE POULBOT ROBOT

The PoulBot (Figure 1) is a track-type mobile robot that has a size of an adult chicken. It is equipped with infrared proximity sensors around the robot, a gyroscope, an accelerometer, a speaker, provides a Bluetooth connectivity and an energy autonomy up to three hours. The robot has an external vision system to track chicks displacement using the overhead camera. To introduce the robot into the animal group we use the filial imprinting mechanism; shortly after hatching we present the moving and vocalizing PoulBot to chicks, afterwards chicks are attracted by the robot and demonstrate the following behavior. Because variations in size, shape and color are well tolerated for imprinting [3], it is not necessary for the robot to look like a hen.

<sup>1</sup> Ecole Polytechnique Fédérale de Lausanne, Switzerland, email: alexey.gribovskiy@epfl.ch

<sup>2</sup> Université Libre de Bruxelles, Belgique, email: jhalloy@ulb.ac.be

The large number of onboard sensors implies that a control system has to be relatively complex; since the main interest and concern of biological researchers lie in proper setting and conducting of animal experiments and not in operating an intricate control interface, our goal is to hide this complexity and to provide a robust and easy to use control system. We use a behavior based controller: the PoulBot is equipped with a number of basic behaviors such as *obstacle avoidance*, *holes avoidance*, *wall following*, *random walking*, *goal following*, etc.; these behaviors can be combined together to form higher level behaviors needed for the current experiments.

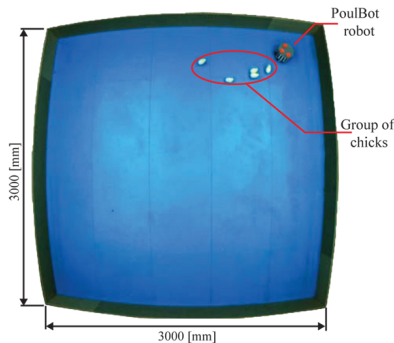


**Figure 2:** The experimental arena with the PoulBot robot and a group of chicks

### 3 EXPERIMENTS

At the moment we run indoor experiments at the Brussels University to study the principles of social aggregation in chick groups. The experimental arena is a flat square 3 by 3 meters surrounded by a wooden wall of 30 cm height (Figure 2). The floor is painted in a blue color to simplify the tracking task.

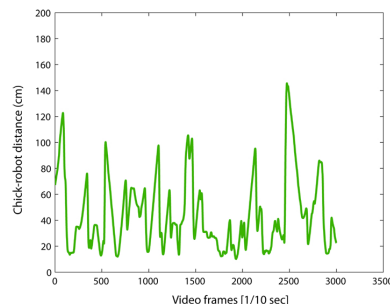


**Figure 3:** The chick-robot distance. The oscillating character of the data indicates the elastic interaction type of the following dynamics between the chick and the robot

Tests are begun by placing a group of chicks (from five to ten individuals) on the arena, where the robot executes a random walking. The chicks are imprinted on the robot, but they are also imprinted on their siblings. We want to study how a chick behavior depends on different cues related to the behavior of the robots and other chicks. We also study the notion of leadership in collective decision making. Chicks have to decide how to deal with their inter-attraction and their attraction towards the surrogate hen. The Figure 3 shows a sample chick-robot distance from one of the experiments; it demonstrates an

elastic interaction dynamics of the chick when it follows the robot, but the exact mathematical model is to be formulated.

### 4 CONCLUSION

In this paper, we have presented current progress on the PoulBot project: we have described the hardware and software of the robot to interact with chicks and have given an example of experiments that we run.

We aim at building artificial systems capable of perceiving specific animal behaviour and to be capable of interacting by appropriate behavioral responses. On the one hand the robots have to reach a sufficient level of cognition for being capable of understanding what animals are doing. On the other hand they have to be capable of emitting appropriate behavioral responses to the animals. These responses are based on known or to be discovered animal cognitive capabilities and behaviors. In other terms the artificial and natural systems has to present some form of mutual relationship, albeit simple at the present stage. Our aim is to build tools that will be used in studies of collective animal behavior that extend classical luring methods. The design of a robotic system is an iterative process closely coupled with a research work on collective animal behavior. Once we have such hybrid systems, we can speculate on the possibility of using the characteristics of each one that are different. The artificial systems could make use of animal cognitive or behavioral capabilities to get information that are inaccessible to it directly by its sensorial capabilities. The natural system could use the capabilities of the artificial system that are foreign to itself like for instance long range communication. This research line is limited by what can be done in artificial intelligence and what we know about animal cognition and behavior. This mirror effect should shed light on their specific limitations and could show where research efforts are necessary in terms of AI and animal intelligence understanding.

### REFERENCES

- [1] A. Goth and C. S. Evans, 'Social responses without early experience: Australian brush-turkey chicks use visual cues to aggregate with conspecifics', *Journal of experimental biology*, **207**, 2199–2208, (2004).
- [2] J. Halloy, G. Sempo, G. Caprari, C. Rivault, M. Asadpour, F. Tache, I. Said, V. Durier, S. Canonge, J. M. Ame, C. Detrain, N. Correll, A. Martinoli, F. Mondada, R. Siegwart, and J.-L. Deneubourg, 'Social integration of robots into groups of cockroaches to control self-organized choices', *Science*, **318**(5853), 1155–1158, (2007).
- [3] Eckhard H. Hess, *Imprinting: Early Experience in the Developmental Psychology*, Van Nostrand Reinhold Company, 1997.
- [4] H. Ishii, M. Ogura, S. Kurisu, A. Komura, A. Takanishi, N. Iida, and H. Kimura, 'Experimental study on task teaching to real rats through interaction with a robotic rat', in *9th International Conference on Simulation of Adaptive Behavior*, (2006).
- [5] J. Knight, 'When robots go wild', *Nature*, **434**, 954–955, (2005).
- [6] P. M. Narins, D. S. Grabul, K. K. Soma, P. Gaucher, and W. Hodl, 'Cross-modal integration in a dart-poison frog', *Proc. Natl. Acad. Sci. USA*, **102**(7), 2425–2429, (2005).
- [7] G. L. Patricelli, S. W. Coleman, and G. Borgia, 'Male satin bowerbirds, *ptilonorhynchus violaceus*, adjust their display intensity in response to female startling: an experiment with robotic females', *Animal Behaviour*, **71** (1), 49–59, (2006).
- [8] A. S. Rundus, D. H. Owings, S. S. Joshi, E. Chinn, and N. Giannini, 'Ground squirrels use an infrared signal to deter rattlesnake predation', *Proc. Natl. Acad. Sci. USA*, **104**(36), (2007).
- [9] R. C. Taylor, B. A. Klein, J. Stein, and M. J. Ryan, 'Faux frogs: multi-modal signalling and the value of robotics in animal behaviour', *Animal Behaviour*, **76**(3), 1089 – 1097, (2008).
- [10] R. Vaughan, N. Sumpter, J. Henderson, A. Frost, and S. Cameron, 'Experiments in automatic flock control', *Robotics and Autonomous Systems*, **31**, 109–117, (2000).

# Context-Specific Vocalizations in Human Infants (Homo sapiens): Can Computer Modelling Help to Bring Further Insight into Categorical Perception?

**Verena Kersken, Juan-Carlos Gómez and Klaus Zuberbühler**

*Centre for Social Learning and Cognitive Evolution, University of St Andrews,  
St Andrews, Scotland, e-mail: vk29@st-andrews.ac.uk*

## **1 Abstract**

Using a methodological approach adapted from field-studies with non-human primates, we observed the vocal behaviour of prelinguistic infants in one of their natural habitat, a nursery environment. We identified a number of vocalizations reliably triggered by specific contexts in the infants and conducted an acoustic analysis of the different call types. Our results show that prelinguistic infants produce different categories of vocalizations. Categorical production has also been observed in other primate species, for example Vervet or Diana monkeys.

Computer modelling could aid the analysis of such data by simulating the perceptual processes at work in the receiver when perceiving and categorizing a call. This would move the way we analyze calls beyond an objective description of their acoustic parameters and more towards a representation of the perceptual processes involved in categorical auditory perception.

## **2 Prelinguistic Vocalizations in Human Infants – Categorical Production?**

Long before human infants produce their first word at about 12 months of age, they produce a variety of sounds that are communicative, for example to express pain, to request objects or actions, to protest about something unpleasant done to them and to point at interesting events or objects in their environment.

In order to examine whether the calls of human infants are functionally distinct, we observed the vocal behaviour of 30 infants between the ages of 7 and 20 months in one of their natural habitats – a nursery environment. We identified the contexts in which vocal behaviour occurred and conducted an analysis of the acoustic features of the vocalizations in five distinct contexts: protests, requests for food or actions, declarative pointing and giving objects to others. A discriminant function analysis was conducted on 10 different acoustic call variables (onset, middle and offset of fundamental frequency; mean and peak intensity, number of harmonics, duration of the call, maximum and minimum fundamental frequency and the number of elements in one call). Overall 192 calls were used for the statistical analysis. The model correctly classified 56.8% of the

calls. We conducted a Monte Carlo simulation to randomize the data by randomly assigning calls categories to the data points and fed it into the model. A chi-squared test was used to compare the overall number of correct and incorrect classifications the model produced. The results revealed a significant difference between the two data sets and that the actual data produced more correct classifications than the randomly generated set (chi squared = 26.99, df = 1,  $p > .0001$ ).

The results suggest that prelinguistic infants produce non-linguistic vocal behaviour selectively in certain contexts. Although this is not a complete description of infant vocal behaviour, it nevertheless suggests some homogenous classes of vocalizations that are systematically related to their production context. Consistency and context-specificity in production provides caregivers and other infants with important cues to what the infant wants or needs. These basic call types therefore function referentially in a comparable way to some vocalizations of non-human primates.

### **3 Context Specificity and Categorical Perception**

The current study applies a methodology developed in the investigation of non-human primate vocal behaviour to human infants. Many non-human primate species produce context-specific calls. Examples include predator-specific alarm calls in Vervet (Cheney, Seyfarth and Marler 1980) and Diana Monkeys (Zuberbühler 2000), both of which have specific calls for aerial or ground predators. Other examples include calls that imply the

presence of different types of food (e.g. Capuchin Monkeys: Pollick, Gouzoules and de Waal 2005; chimpanzees: Slocombe and Zuberbühler 2005).

Playback studies have demonstrated that some of these calls are meaningful to other primates and that these calls are perceived categorically (Cheney, Seyfarth and Marler 1980, Zuberbühler 2000). Categorical perception has been studied extensively in humans, particularly with regards to how we perceive spoken language and its various elements (e.g. Brauer and Friederici 2007). In categorical perception, there are psychological mechanisms at work which transform the actual perceived sound to a virtual sound that is similar to the nearest category prototype - so it can be classified into one category despite not being a perfect match (Livingston et al 1998). These mechanisms might ameliorate particular properties of the heard sound, such as a particular pitch or frequency pattern, and therefore allow the recipient to classify the sound into one particular category more readily. MAYBE TOO TEXTBOOK LIKE

### **4 Challenges for Computer Modelling**

If categorical perception in non-human primates also works in this particular way, then the statistical analyses we are using at the moment are unable to produce meaningful results because it is not clear which parameters are relevant for the understanding and categorization of these calls. The call analyses primatologists are using measure objective parameters, such as frequencies, duration, harmonics or intensities to describe calls and

compare them against each other to identify potential differences between classes of calls. However, although this method seems valid for the objective description of calls, the choice of parameters is often arbitrary and, more importantly, might not be a valid reflection of the processes going on in the receiver.

Computer modelling could help to expand and improve the method of analysis currently used with research into primate vocal production and perception and infant prelinguistic vocalisations. Computer modelling could offer a form of data analysis that more closely resembles the actual perceptual processes that are at work in the receiver. Projects so far have mainly considered the area of language and speech perception. however rarely the non-linguistic sounds produced by non-human primates or prelinguistic infants.

Human infants are a particularly interesting example because they interact with a linguistic community and are able to understand speech, but are yet unable to produce language. It could for example identify salient features of calls and model understanding and categorisation in the receiver. Investigating call classes using model clusters could therefore be a more appropriate way of data analysis. Furthermore, the wealth of data collected in this topic can provide a rich database to model brain and perceptual processes that are working in the perceiver.

### Acknowledgements

Thanks to Acorn and WonderYears Nurseries, Eric Bowman and the Leverhulme Trust

### References

Brauer, J. and Friederici, A.D. (2007): “*Functional neural networks of semantic and syntactic processes in the developing brain*”, *Journal of Cognitive Neuroscience*, Vol.19, No. 10, pp. 1609-1623

Cheney, D.L., Seyfarth, R.M. and Marler, P. (1980): “*Vervet monkey alarm calls: Semantic communication in a free-ranging primate*”, *Animal Behaviour*, Vol. 28, pp. 1070-1094

Livingston, K.R., Andrews, J.K. and Harnard, S. (1998): “*Categorical Perception Effects Induced by Category Learning*”, *Journal of Experimental Psychology: Learning, Memory and Cognition*, Vol. 24, No.3, pp. 732 - 753

Pollick, A.S., Gouzoules, H. and de Waal, F.M.B. (2005): “*Audience effects on food calls in captive brown capuchin monkeys, *Cebus apella**”, *Animal Behaviour*, Vol. 70, Issue 6, pp. 1273-1281

Slocumbe, K.E. and Zuberbühler, K. (2005): “*Functionally Referential Communication in a Chimpanzee*”, *Current Biology*, Vol. 15, Issue 19, pp. 1779-1784

Zuberbühler, K. (2000): “*Referential Labelling in Diana Monkeys*”, *Animal Behaviour*, Vol. 59, Issue 5, pp. 917-927

# Generation of Biological Motion Patterns and Their Emergent Transitions with a Trainable Neural Network Oscillator

Emilian Lalev<sup>1</sup>

**Abstract.** Biological Motion perception is the ability to perceive gestalts of moving human or animal figures from two-dimensional projections of their joints. These projections are very salient stimuli. A neural network system for generation of such sequences and transitions between them is described. The network can produce several sequential motion patterns learned from human motion capture data. Realistic transitions between them can then be emergently generated while switching between modes of oscillation (e.g., from one type of gait to another). This system might be further explored as a cognitive model for human motion generation showing that motions can be stored in exactly the same structure.

## 1 Introduction

Biological Motion (BM) perception was first investigated by [8, 9]. It refers to the ability to perceive moving human or animal figures only given the projections of their joints on a 2D plane (the so-called point-light displays) [8]. Human movements contain many degrees of freedom and include both rigid and nonrigid motion which makes them very complex. BM carries rich information about the perceived object: First, it works for identification of animals and other people from their surrounding [14]. The observed body kinematic structure is revealed and the identity of the living being is classified [2]. BM also reveals the particular type of action being performed, such as walking, running or dancing. And finally, perception of BM can even help define qualitative characteristics of the agent such as gender, weight, mood, etc. [4].

The phenomena related to BM perception and production are approached by several computational and statistical methods. The so called "morphable models" [6, 10] have been applied for the purposes of motion editing [1], re-targeting motion from one character to another [7], tracking a human figure from video data [11], or recognizing type and style of gait patterns [6]. Fourier techniques [13, 3] are based on analyses of sinusoidal oscillations and are applied for periodic motions, such as locomotion patterns.

In addition to the former approaches, an artificial neural network system was designed to process BM [12]. In the system, periodic point-light displays are approximated as combinations of complex sinusoidal functions. The advantage of such a system relative to similar existing ones is that generalization properties of neural networks seem able to emergently solve particular problems of motion generation and transitions, thus reducing computational demands and efforts for the user.

<sup>1</sup> Central and Eastern European Center for Cognitive Science, New Bulgarian University, Sofia Bulgaria, 1618 Montevideo Str., email: elalev@cogs.nbu.bg

## 2 System description

An Elman recurrent network [5] and one feed-forward network are the parts of the system. The recurrent network (BM Generator) is used as a trainable oscillator. One single network generates each of several learned point-light displays and transitions. It is optimized with the feed-forward network (Transition Filter, representing the physical restrictions of the moving body) to ensure the smoothness of transitions. The work of the Transition Filter is to try to associate only known body postures from novel and noisy frames initially obtained from the Generator during transitions.

### 2.1 Networks

The BM Generator (Figure 1) has 300 hidden units. The Transition Filter (also included in Figure 1) has 150 hidden units. Both networks have tanh-sigmoid transfer functions.

### 2.2 Inputs and Targets

As training data for the system, human motion capture data of several motions is used. 15 markers are chosen/computed from the original recorded markers [13]. The coordinate system that is used is object-centered. The number of frames for a particular motion depends on the speed of motion execution. The temporal resolution is 24 frames per second.

From every motion capture record, one locomotion cycle is extracted such that the dynamic pattern can be repeated infinitely.

The input training data for the Generator is composed of the classification vector (or the label) of a given motion repeated constantly in time, quite similar to bias connections attached to the input nodes (see Figure 1).

The training inputs and targets for the Transition Filter are identical as it learns to auto-associate each known frame of the trained motions. During the work of the system, the Transition Filter takes as input the output from the Generator (see Figure 1).

### 2.3 Training

An algorithm for adaptive training goal changes the goal according to the current performance. The networks are trained to fit 2 different walking patterns, and one running pattern as the mean squared error does not exceed 0.001.

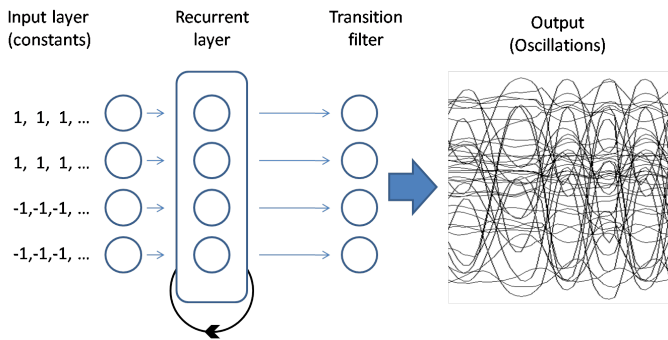


**Figure 1.** Architecture of the BM Generator and its inputs and targets with the transitions filter.

### 3 Results from the BM Generator

1. It learns to oscillate infinitely each of several dynamical patterns.
2. The Generator can perform switching between different BM sequences when the input is changed accordingly. Depending on the exact frame that switching happens, different emergent transitions may be observed.
3. Despite the fact that the BM generator is not trained to any transitions, it generates them quite plausibly in most cases. One such example is accelerating during switching from walking to running. It is shown in the short animation *transitions1.gif* at <http://sites.google.com/site/emilianzahariev/lealev/upload/>
4. Transitions are additionally smoothed and restricted to realistic transformations by the Transition Filter network (Figure 1) which reproduces the coordinates of each motion frame. Filtered transitions are qualitatively smoother and more realistic than the ones that come directly from the network's output. A short animation *transitionsFilteredAndRaw.gif* showing this is available at <http://sites.google.com/site/emilianzahariev/lealev/upload/>

Smoothness of transitions can also be quantitatively measured by the mean absolute value of differentiated oscillations (Figure 2). In the figure, a transition is initiated at frame 21 and ends about frame 40. Obviously, the usage of the auto-associator (associator1) smoothes the sharp changes in this period. Applying it once again (associator2), smoothing of the border between the motions becomes even better.

### 4 Conclusion

I showed implementation of recurrent neural networks in processing of real human motion capture data. Recurrent networks can be utilized as trainable oscillators of different body motions. Switching between oscillations is smooth and very realistic. Such smooth transitions are due to the recurrent activations in the network remaining from the previous oscillating regime. They work as initial context for the new oscillation.

Networks learn dynamics in BM as approximated functions. They develop several dynamical attractors, resulting in oscillations representing separate locomotion patterns. Maybe this system has common grounds with the biological mechanisms for oscillatory motions like human and animal gaits. The basic inference would be that a particular and exactly the same brain structure may be responsible for several different command tasks.

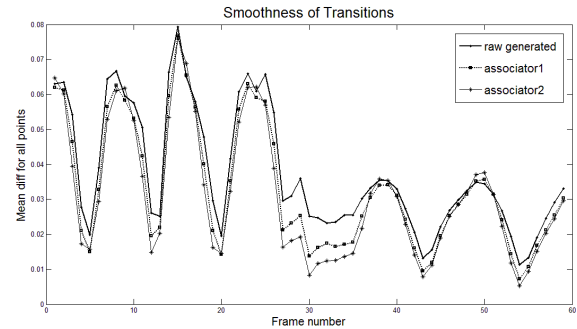


**Figure 2.** Differentiated coordinates' trajectories as a measure for the smoothness of transitions from the BM Generator - raw or processed once or twice by the transition filter.

### ACKNOWLEDGEMENTS

I would like to acknowledge Prof. Dr. Nikolaus Troje, head of Biomotion Lab, Queen's University, Kingston, Ontario, Canada for providing me with human motion capture data.

### REFERENCES

- [1] A. Bruderlin and L. Williams, 'Motion signal processing', in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles*, pp. 97–100, (1995).
- [2] J. E. Cutting and L. T. Kozlowski, 'Recognizing friends by their walk: Gait perception without familiarity cues', *Bulletin of the Psychonomic Society*, **9**(5), 353–356, (1977).
- [3] J. Davis, A. Bobick, and W. Richards, 'Categorical representation and recognition of oscillatory motion patterns', in *IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, SC*, pp. 628–635, (2000).
- [4] W. H. Dittich, T. Troscianko, S. E. G. Lea, and D. Morgan, 'Perception of emotion from dynamic point-light displays represented in dance', *Perception*, **25**(6), 727–738, (1996).
- [5] J.L. Elman, 'Finding structure in time', *Cognitive Science*, **14**, 179–211, (1990).
- [6] M. A. Giese and T. Poggio, 'Morphable models for the analysis and synthesis of complex motion patterns', *International Journal of Computer Vision*, **38**, 59–73, (2000).
- [7] M. Gleicher, 'Retargetting motion to new characters', in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, Orlando*, pp. 32–42, (1998).
- [8] G. Johansson, 'Visual perception of biological motion and a model for its analysis', *Perception and Psychophysics*, **14**, 201–211, (1973).
- [9] G. Johansson, 'Spatio-temporal differentiation and integration on visual motion perception', *Psychological Research*, **38**, 379–393, (1976).
- [10] M. J. Jones and T. Poggio, 'Multidimensional morphable models: A framework for representing and matching object classes', *International Journal of Computer Vision*, **29**, 107–131, (1999).
- [11] S. X. Ju, M. J. Black, and Y. Yacoob, 'Cardboard people: A parameterized model of articulated image motion', in *Proceedings of the International Conference on Face and Gesture, Vermont*, (1996).
- [12] E. Lalev, 'Connectionist framework for processing of biological motion sequences and their transitions', in *Proceedings of the 20th Irish Conference on Artificial Intelligence and Cognitive Science AICS2009*, eds., John Dunnion Lorcan Coyle and Jill Freyne, pp. 198–209, (2009).
- [13] N. F. Troje, 'Decomposing biological motion: A framework for analysis and synthesis of human gait patterns', *Journal of Vision*, **2**, 371–387, (2002).
- [14] N. F. Troje and C. Westhoff, 'The inversion effect in biological motion perception: Evidence for a 'life detector'?', *Current Biology*, **16**(8), 821–824, (2006).

# Innate movement strategies for view-based homing

Thomas Parslow and Andrew Philippides and Paul Graham<sup>1</sup>

**Abstract.** Here we investigate an innate movement strategy for view-based homing using snapshots containing only panoramic skyline height information. The method is developed in simulation but implementation in a real robot is planned.

## 1 INTRODUCTION

Behavioural experiments have shown that insects can return to a location using only a remembered view or snapshot taken at that location [1, 5]. By comparing the sum-squared-difference between a goal image and images taken around that goal, Zeil and colleagues showed that the difference increases smoothly with distance in natural environments [7]. Thus a smooth gradient exists in image space in natural environments which an agent can follow to return home. An example of this gradient can be seen in Figure 1. The task is then to sense this gradient and move down it to find the goal. The gradient is unlikely to ever be perfect and will often contain local minima so a successful strategy must be capable of moving over these depressions without getting stuck. Several algorithms have been developed that implement this method of homing (e.g [7, 3, 6]).

Recent experiments have shown that skyline is important for navigation in certain types of ants and bees [2, 4]. The skyline provides a simplification of the scene which is fairly robust over multiple journeys and in differing lighting conditions. The skyline is also an easy feature to extract. Here we investigate an innate movement strategy for navigation down the view-based gradient given by the skyline, without the use of explicit sampling movements. Specifically, the agent moves in a sinusoidal wiggling path as illustrated in Figure 1 with rate of turning modulated by the recently observed change in skyline with regards to the goal snapshot.

## 2 METHODS

A simple route following task is used to test our proposed method. The agent is placed at a point in a simulated environment consisting of columns of different heights and widths. The task simulates route following so the starting positions are placed so that the agent is roughly facing the goal. This simplifies the task somewhat and means that the harder cases such as when the agent is facing the opposite direction are removed. The agent moves in a sinusoidal fashion which is implemented by oscillating the orientation of the agent between  $\pm 45$  degrees while moving it forwards at a constant speed.

The only information available is the output of a 360 degree panoramic sensor which gives a heightmap consisting of the perceived height of objects against the skyline. The sensor is held in a fixed orientation relative to the world. At regular intervals a new

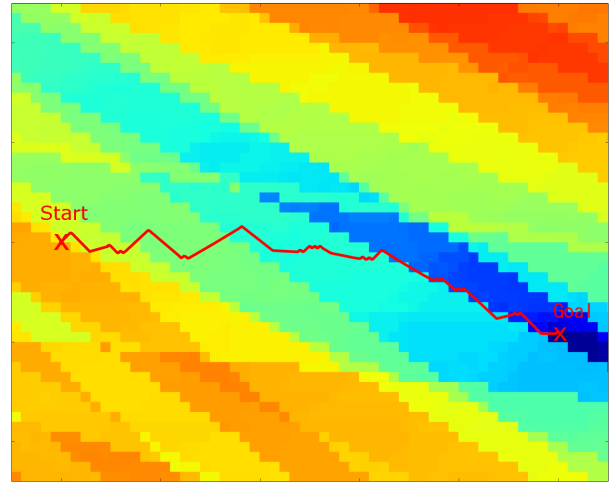


**Figure 1.** Background colours show the skyline view-based gradient and the red line shows the path of single fairly typical run of the simulated agent.

heightmap is retrieved from the skyline sensor. As input to the control algorithm the skyline heightmap is reduced to a single error variable by taking the root mean square difference (RMS) between the skyline heights at the current position and a ‘snapshot’ taken at the goal location and orientation. This single variable is used to modulate sinusoidal path of the agent, by setting the speed of oscillation. Thus when movement is in a correct direction the speed of turn slows and when the direction is less good it speeds up, thus the agent spends

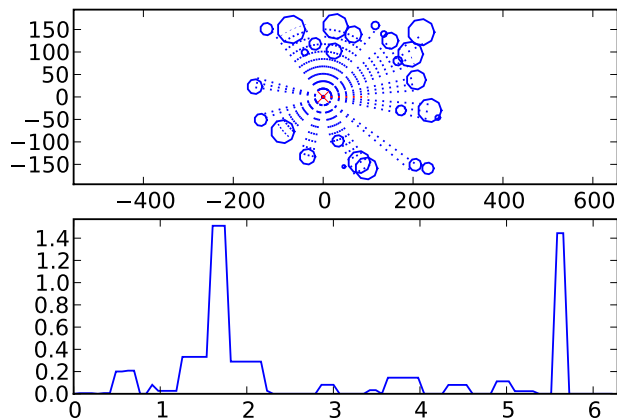


**Figure 2.** The simulated world with 360 panoramic horizon view.

<sup>1</sup> University of Sussex, United Kingdom, email: {twp21, andrewop, paulgr}@sussex.ac.uk

much more time moving in “good” directions. A running average of the last few samples is used as a comparison to help smooth out noise.

### 3 RESULTS AND DISCUSSION

Early results show the strategy resulting in fairly robust homing behaviour when the snapshot orientation is fixed and modulation of the turning speed is used. The simulated agent is able to find a goal location with a high degree of accuracy in most runs when the angle to goal position is within  $20^\circ$  of the agent’s facing orientation. Figure 1 shows a typical run of the simulated agent. The strategy continues to function well in the presence of a fair amount of motor noise and visual noise.

These results from simulation show promise for our next goal of implementing the methods on a mobile robot platform capable of outdoors operation. This will provide useful information on the suitability of the proposed methods under noisy real world conditions as would be experienced by insects and so shed light on their plausibility. However, there are several extensions that need to be investigated before a robotic implementation can be realised.

Robust skyline extraction is the first thing needed. The wheeled robotic platform makes use of a 360 panoramic camera. This provides an image that can be unwrapped to a 360 degree panoramic image (Figure 3). Skyline extraction can be implemented very simply, a simple threshold on the blue channel of the image detects the point at which sky becomes non-sky.

Secondly, in the simulation the snapshot view orientation was fixed throughout the run while the direction of movement oscillated. This fixed view orientation version could be considered “bee-like” behaviour, while an “ant-like” version would result in the orientation changing with the movement direction. The “ant-like” is more appropriate for implementation on the wheeled robot as the fixed view orientation requires absolute heading information that is not always reliably available in the real world.

While the fixed view orientation strategy was found to be more successful in the simulation than preliminary tests with the alternative, it is hoped that the situation may reverse on the real robot. The results in simulation, with its overly simple and featureless columns, may be due to the poor rotational gradient. Initial tests in the real world show a much stronger rotational gradient which is expected make the second strategy successful.

### ACKNOWLEDGEMENTS

TP is funded by an EPSRC studentship.

### REFERENCES

- [1] B. A. Cartwright and T. S. Collett, ‘Landmark learning in bees’, *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, **151**(4), 521543, (1983).
- [2] Paul Graham and Ken Cheng, ‘Ants use the panoramic skyline as a visual cue during navigation’, *Current Biology*, **19**(20), R935–R937, (November 2009).
- [3] R. Möller and A. Vardy, ‘Local visual homing by matched-filter descent in image distances’, *Biological cybernetics*, **95**(5), 413430, (2006).
- [4] W.F. Towne and H. Moscrip, ‘The connection between landscapes and the solar ephemeris in honeybees’, *Journal of Experimental Biology*, **211**(23), 3729, (2008).
- [5] A. Vardy and R. Möller, ‘Biologically plausible visual homing methods based on optical flow techniques’, *Connection Science*, **17**(1), 4790, (2005).

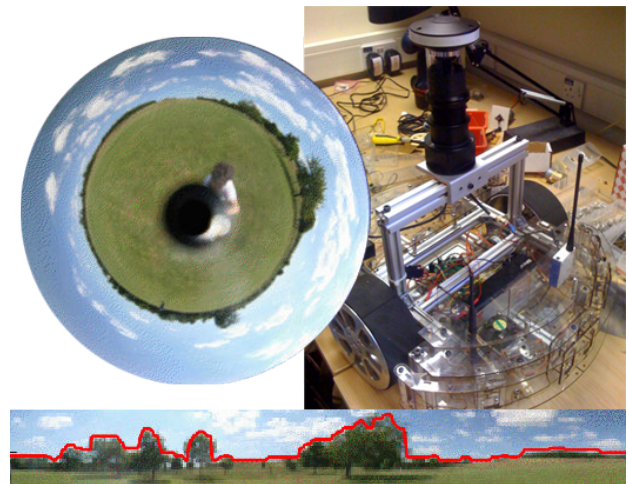


**Figure 3.** A wheeled robotic platform with a 360 panoramic camera mounted along with an image from the camera and the unwrapped version with skyline height.

- [6] M. Zampoglou, M. Szenher, and B. Webb, ‘Adaptation of controllers for image-based homing’, *Adaptive Behavior*, **14**(4), 381, (2006).
- [7] J. Zeil, M. I. Hofmann, and J. S. Chahl, ‘Catchment areas of panoramic snapshots in outdoor scenes’, *Journal of the Optical society of America A*, **20**(3), 450469, (2003).

# Corvid: A Versatile Platform for Exploring Mobile Manipulation

Michael Zillich<sup>1</sup> and Michael Baumann and Wolfgang Knefel and Christoph Längauer

There is an abundance of different robot platforms used in various fields of robotics. Some are custom made, others are more or less standardised off-the-shelf products. In fields such as epigenetic robotics or swarm robotics these are typically small inexpensive platforms of limited complexity (“ton on wheels” design) such as the widely used K-Team Khepera [10], EPFL e-puck [7], Eddy [1] or Surveyor SRV-1 Blackfin [15] to name a few. These platforms are well suited to operate in specially prepared “arenas” (enclosures with various markers or obstacles, mazes etc.) and to perform experiments on navigation and obstacle avoidance. Due to their small size and simplicity however they are not so well suited to perform more complex tasks including manipulation (though the Khepera can be equipped with a 2 DOF gripper).

More complex platforms like the late Sony Aibo, or more recently humanoids like iCub [14], the Fujitsu hoap series or the Aldebaran Nao [5] offer many degrees of freedom and a sturdy design. However often only a subset of the DOFs is used in learning experiments, e. g. with a sitting humanoid where only the arms and perhaps torso are moved.

Wheeled platforms such as Stanford’s STAIR [3], DLR’s Justin [6] or UMass’s uBot [4] offer effective mobility and good manipulation skills. Their size, complexity and also price however limit their applicability in autonomous learning: errors made by such a platform are expensive and dangerous.

A MobileRobots Pioneer 3-DX with optional 7 DOF arm seems the best choice so far, but is still on the heavy, slightly dangerous and expensive side (compared to lightweights like the Khepera). Also the rather high mounting of the small arm limits its workspace at floor level.

In summary, none of the above fulfilled our requirements for performing experiments in learning mobile manipulation:

- inexpensive to purchase and maintain
- safe to operate, to allow use in student projects
- able to carry an arm
- sufficient onboard processing power
- WLAN capabilities to enable offloading compute-intensive tasks to a PC
- possibility to connect standard peripherals like webcams via USB
- able to traverse “rough terrain” like door thresholds and cables, i. e. able to navigate on the floor in typical labs rather than in specific “arenas”

To fill this gap we present the robot platform Corvid. The

name Corvid derives from the configuration of the gripper and camera (an eye in hand setup) which aims to mimic corvids (ravens, crows, jays) which are adept manipulators where similarly the eyes are in a fixed relation to the beak.

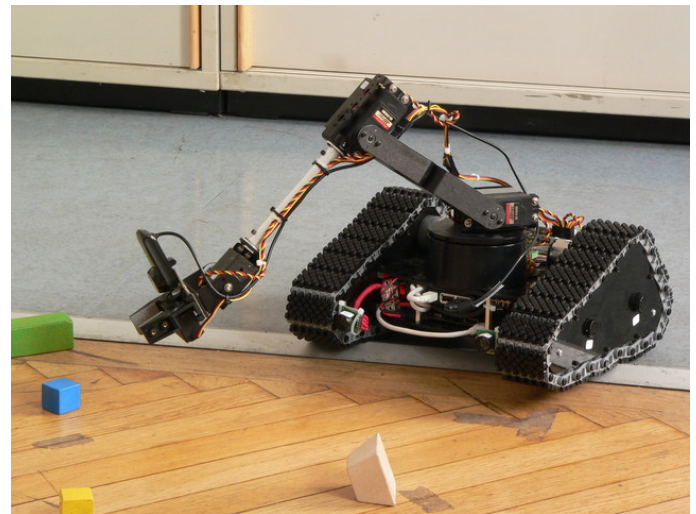


Figure 1. Corvid robot platform

Nature basically developed two concepts for manipulation. The first one involves eyes and independently moving manipulators, as in hominids (various apes and of course humans) but also sea otters or octopi. The second one involves eyes and manipulator (beak, mandibles) in a fixed relation, as in birds (of which corvids are amongst the most skilled) and various insects. Note that birds often also use their feet very adeptly, typically in conjunction with the beak. So strictly speaking they belong in both categories. It seems likely that the added complexity of independently moving manipulators fostered the development of advanced cognitive skills in such species. From that point of view our robot platform is certainly more insect-like than bird-like. But our point here is not to define strict categories (which is often more a hindrance than a help anyway) but to indicate where we place our robot design with respect to biological examples. And a bird with a flexible neck looking around and pecking and picking things seems to be the closest biological relative.

On an orthogonal axis we can distinguish manipulators composed of jointed limbs vs. highly flexible manipulators (such as the tentacles of octopi, an elephants trunk, or an ant-

<sup>1</sup> Vienna University of Technology, Vienna, Austria, email: zillich@acin.tuwien.ac.at

eaters tongue). The latter would seem to rely more heavily on fine-grained tactile information rather than visual feedback, and as such is very difficult to achieve with current tactile sensing technology.

While independently moving manipulators allow greater flexibility, and are mimicked in humanoid robots, we opted for the fixed manipulator-eye arrangement for various reasons. The smaller number of degrees of freedom generally results in smaller (and thus hopefully more manageable) learning problems. Also from a pragmatic engineering perspective it is easier if the manipulator can not obscure the view of the camera, especially as the gripper approaches an object. Humanoid robots either have a rather limited workspace right in front of the chest with arms coming in from the sides, or would require a long, flexible neck otherwise. Moreover mounting the camera on the arm provides an active camera without the need for an extra pan/tilt unit.

**Hardware** The platform basically follows the ton on wheels design, where however the wheels are replaced with threads. The large contact area of the threads improves stability compared with only three contact points in case of the typical two powered wheels plus caster wheel. The threads also allow mounting low obstacles like door steps of up to about 3 cm height. This design is popular in field and rescue robotics, e. g. the Mesa Robotics Matilda [12] or iRobot Pack-Bot [9] series. These platforms are built for extreme durability which is reflected in their steep price tag.

We chose the inexpensive Lynxmotion TriTrack chassis [11] as a basis and added a Lynxmotion AL5D arm with 4 degrees of freedom plus gripper. A fifth degree of freedom can be added via an optional wrist rotate joint. The arm has a forward reach of about 260 mm (see Figure 2) and can carry loads of up to 300 g. As can be expected from an arm powered by RC servos, due to joint backlash accuracy is rather low with maybe  $\pm 5$  mm in fully extended position. The overall dimensions of the platform are  $L \times W \times H = 270 \times 280 \times 560$  mm (arm fully upright) with a weight of 3 kg.

A 2600 mAh Ni-MH battery provides 12 V for the drive motors and a power regulator provides additional 5 V for the RC servos of the arm and the electronics. The two 12 V DC drive motors are controlled by a Dimension Engineering Sabertooth 2x5 dual motor controller. The RC servos of the arm are controlled by a Lynxmotion SSC-32 Servo Controller.

The main robot controller is a gumstix Overo Fire board [8] with a Texas Instruments OMAP 3530 CPU (an ARM derivative) running at 600 MHz, carrying 256 MB RAM, 256 onboard flash and an additional 2 GB of microSD flash memory. It also offers a DSP and 3D graphics acceleration. The board furthermore provides USB 2.0, WLAN and bluetooth, HDMI video output, Audio I/O, I2C, A/D and PWM lines. A powered 4-Port USB hub allows connection of additional peripherals, like a webcam.

Regarding sensors, the robot is equipped with 8 Devantech SRF-02 ultrasonic range sensors for measuring distances from 15 cm up to 6 m, connected via I2C bus. A Logitech Quick-Cam Pro USB webcam with a high quality auto focus Zeiss lens is mounted on the gripper and provides images of up to  $1600 \times 1200$  Pixel at frame rates of up to 30 Hz.

**Software** The gumstix board runs Linux (kernel 2.6.29) and thus supports almost any peripherals You can think of.

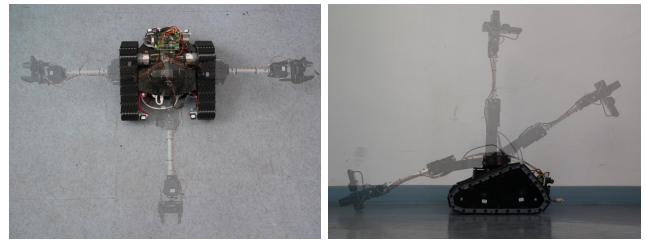


**Figure 2.** Manipulator workspace

Development is based on the OpenEmbedded [13] development environment.

Corvid offers a driver for the Player project [2] that implements the position2d (moving about in 2D), sonar, camera, actarray (actuator array - the arm) and gripper interfaces. Support for the limb interface (inverse kinematic control of the arm) is planned. The player server running on the platform can be accessed via WLAN.

**Conclusion** The platform we presented is an inexpensive option to start exploring mobile manipulation. The complete platform costs around 1500 EUR (excluding tax and shipping), which is mainly composed of 250 EUR for the chassis, 500 EUR for the arm, 350 EUR for the controller board and the rest for batteries, sensors and electronics.

The emphasis lies on maintainability and safety, to allow longer term unattended experiments and also operation by students without the worry of dangerous and expensive accidents. And if things do go wrong, single servos or DC motors can be replaced for 30 to 40 EUR.

The robot controller provides sufficient processing power to run software of medium complexity including image processing onboard. Its rich connectivity offers the possibility to add own sensors (for example IR distance sensors via I2C or even a Hokuyo laser range finder via USB) or further actuators like RC servos.

We will provide a webpage which summarises all parts and instructions at [www.acin.tuwien.ac.at/1/research/v4r](http://www.acin.tuwien.ac.at/1/research/v4r).

**Acknowledgements** This work was supported by the City of Vienna University Jubilee Trust project “LEMMA” (H-1773/2008) and EU FP7 project “CogX” (ICT-215181-CogX).

## REFERENCES

- [1] [metropolis.sci.univr.it/eddy](http://metropolis.sci.univr.it/eddy).
- [2] [playerstage.sourceforge.net](http://playerstage.sourceforge.net).
- [3] [stair.stanford.edu](http://stair.stanford.edu).
- [4] [www-robotics.cs.umass.edu/Robots/UBot](http://www-robotics.cs.umass.edu/Robots/UBot).
- [5] [www.aldebaran-robotics.com](http://www.aldebaran-robotics.com).
- [6] [www.dlr.de/rm/en/desktopdefault.aspx/tabid-5471](http://www.dlr.de/rm/en/desktopdefault.aspx/tabid-5471).
- [7] [www.e-puck.org](http://www.e-puck.org).
- [8] [www.gumstix.com](http://www.gumstix.com).
- [9] [www.irobot.com](http://www.irobot.com).
- [10] [www.k-team.com](http://www.k-team.com).
- [11] [www.lynxmotion.com](http://www.lynxmotion.com).
- [12] [www.mesa-robotics.com](http://www.mesa-robotics.com).
- [13] [www.openembedded.org](http://www.openembedded.org).
- [14] [www.robotcub.org](http://www.robotcub.org).
- [15] [www.surveyor.com/blackfin](http://www.surveyor.com/blackfin).

# Modelising H-CogAff global alarm system improves robot performance

Mary Felkin and Yves Kodratoff<sup>1</sup>

**Abstract.** We developed a system which purpose is to obtain a robot able to emulate the strategies used by a subject facing a problem-solving task. We have been able to solve this problem within a strongly constrained setting in which the subject's strategy can be induced. Our solution encompasses the solution of another problem, namely how to close the raw-data/trained-system/raw-data loop. An important aspect of inducing subject strategies was becoming able to recognise changes of strategy. The H-CogAff from which our control architecture is inspired (though we do not yet attempt to completely emulate it) provides a handy trigger for strategy changes: the global alarm system. We show that implementing it improves the robot's performance.

## 1 Introduction

There are many abstract definitions of human intelligence and not enough engineering blueprints of it. Paradoxically, working models abound all around us: human beings. So we teamed up with psychologists to study and emulate subjects in a problem-solving context. Though the different relevant behaviours of human beings in a limited environment and with a fixed goal can be identified and simulated, the strategies which go on in the higher levels of the "control architecture"<sup>2</sup> which structure these behaviours are harder to induce: doing so requires a careful step-by-step approach. To guide us in exploring the as yet incompletely chartered territory of the subject "control architecture" we chose the H-CogAff map. This turned out successfully so we are now taking a step further and tying our model more closely to H-CogAff by implementing a global alarm system.

From an AI/Robotics standpoint, when a mechanism which mimics a theoretical human thought process improves the performance of the robot, this validates the theory.

## 2 Related work

A large amount of work has been done in the field of robot learning by imitation, a relatively new (about twenty years old) field of research, see for example [5], [10] and [19]. This field takes inspiration from a wide range of disciplines, including psychology, biology, neurobiology, etc. [2], [4], [9] and [6]. An example among others of the necessary multidisciplinary is [3] who propose a mathematical solution to the correspondence problem, which originally comes from animal psychology: they formalise the correspondences by giving mapping matrices to link agents with different morphologies. Other research papers present architectures which are less biomimetic, for

example [7] who present an architecture for extracting the relevant features of a given task and then generalise the acquired knowledge to other contexts. They demonstrated the effectiveness of their architecture by implementing it on a humanoid robot learning to reproduce the gestures of a human teacher, not all architectures are inspired by cognitive sciences, our approach is just one among many. A formal definition of plan recognition can be found in [15]. Close to our work but using a different mathematical framework are Bayesian methods: [24] focuses on the higher levels given a set of basic robotic actions and [20] on real-time Bayesian learning. [14] compare several approaches and discuss imitation learning in virtual reality.

## 3 The raw-data/trained-system/raw-data loop

This loop, which to the best of our knowledge we were the first to close, enables for example a robot to learn from another robot. When domestic androids will reach the mass production stage few users will be happy to train their robot from scratch. Generalising human strategies from many examples of tasks such as clearing the table after a meal in different settings will enable the designers to abstract these strategies from their particular context. A simulated robot could then enact them (at first for testing purposes) in a much wider variety of circumstances, including unlikely ones (for example in cases where the dish washer would be in the cellar). The simulated robot could clear tables in a million different simulated houses. So subsets of cases in which the robot fails could be identified and a new strategy learnt for these cases.

Once the robot's behaviour becomes fully adequate for the task, its recorded actions would provide as much learning data for subsequent robots as would be required for robust generalisation: Abstracted strategies are also abstracted from body built because our method oversteps the correspondence problem<sup>3</sup> so the same strategy could be implemented by many different androids.

This allows incremental additions to the robot's capacities because as we will see logs can be combined. The android can learn by imitating its user that the blue glasses go into the glass cabinet and smoothly add this knowledge to its previous strategy.

Our raw-data/trained-system/raw-data loop is a loop starting with the behaviours of several subjects and their analysis and interpretation in terms of human observable actions. This leads to the definition of the strategies used by the subjects (including inefficient ones), the interpretation of the human observable actions in terms of movements of the robot and to the definition of what is a "robot

<sup>1</sup> CADIA, University of Reykjavik, Iceland and LRI, Université Paris-Sud, France

<sup>2</sup> Strategies are considered to be a subset of meta-management processes.

<sup>3</sup> In robotics, the "correspondence problems" refers to the problems which arise from the fact that no (existing) robot is built exactly like a human. A humanoid robot may have a morphology and physical capacities which are quite different from these of its instructor [3]. In our example the simulated "teaching robot" and the "learning robots" may all be built in different ways

strategy” in terms of human strategies. Our loop is closed with a programming language enabling us to implement these robot strategies in new settings, either by hand or automatically from the log of the subjects’ actions in the original settings. When the induced strategies are enacted, they become observable in the same way as the subject’s strategies were observable at the beginning of the loop.

## 4 Experimental settings

Our final goal was not to replicate the behaviour of individual volunteers but to “understand” their underlying strategy and to become able to reproduce it in new surroundings.

In a sequence of psychological experiments, blindfolded human volunteers explored a maze<sup>4</sup> in search of a “treasure” (a bottle of water) and, while doing so, they expressed their search strategy<sup>5</sup> by sequences of perception-actions pairs, which were recorded. Perception here was limited to touch, which could be observed on the videos. Actions were limited to moving in the maze, touching objects and picking up the treasure, so these could also be observed. Fig. 1 illustrates the process of capturing the volunteers movements in our simulator and then recording the databases of observable actions.

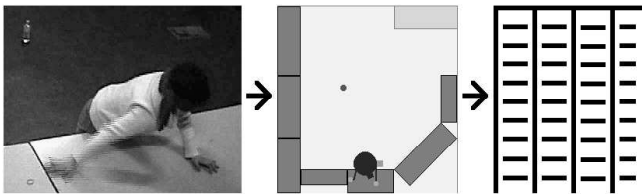


Figure 1. From videos to databases

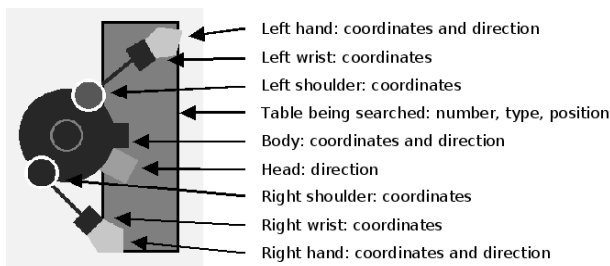


Figure 2. Some of the observable variables

The psychologist [13] and the mixed team [23] showed that the volunteers had several different goals which they combined through some thought process akin to multi-criteria optimisation to mentally construct and evaluate their behaviours. On top of their given goal, finding the treasure, their most often used strategies included the goals of not getting lost, of not exploring the same place twice, of not bumping into obstacles, etc. We performed a detailed analysis,

<sup>4</sup> The mazes were not virtual, they were built with rows of tables and sometimes cupboards and radiators in a large schoolroom.

<sup>5</sup> Our use of the expression “search strategy” here does not imply the volunteers were searching according to an explicit plan. Randomly searching through the maze is also a “search strategy”, and so is “not searching at all”.

including a digitalisation, of the videos showing the behaviour of 10 of these volunteers, called G1.1, G1.2, G1.3, G3.1, G3.2, G4.1, G4.2, G4.3, G7.1 and G7.2 in the following. We thus could run a close<sup>6</sup> replicate of their behaviour in our system and analyse it.

## 5 A step-by-step analysis

Automatically extracting from a database the strategies used by subjects in a problem-solving situation takes more than a good pre-processing and then running the database through the appropriate data mining algorithm. To go from the database of observables to the strategies, we had to define a middle ground. Fig. 3 models the human’s cognitive processes as a very simplified version of the H-CogAff (Human Cognitive Affects, [21]) model, and superimposes our definitions.

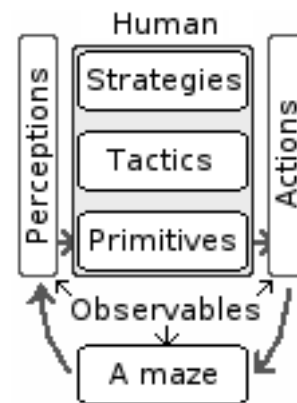


Figure 3. From observables to strategies

The raw data contained in the databases, called the observables, are indicated in fig. 2. They are the basic facts such as the position of the person in the maze at a given time step, the position of his/her hands, etc. 50 observables were recorded every quarter of a second in what we call the “log files” or “databases”. Each maze also has a static description indicating the position of the obstacles, of the treasure, etc. Primitives are combinations of observables, and sometimes of observables and static maze descriptors. We call “tactic” a combination of observables and primitives linked by a common goal or sub-goal. We defined four tactics:

- the goal-related treasure hunting tactic, called the “search tactic”,
- the tactic used by the volunteer to cope with the fact that he or she has to move around blindfolded, called the “moving tactic”,
- the tactic causing the behaviour of the volunteer encountering an obstacle, which has a mixed purpose of treasure hunting and spatial orientation, called the “obstacle following tactic”
- the personal safety tactic called the “obstacle detection tactic”.

In [11] we detail why and how this particular decomposition was chosen. The combined effect of enacting each of the four types of tactics is a strategy.

Fig. 4 shows the path we followed in this work: First a bottom-up generalisation, in several steps, which started with the log file recording the movements of the subject and was achieved with the help of machine learning algorithms. Then the top-down implementation of

<sup>6</sup> Our replicates are only “close” because the recordings are noisy. We believe this noise contributes to the robustness of the ensuing generalisations but we did not test this hypothesis.

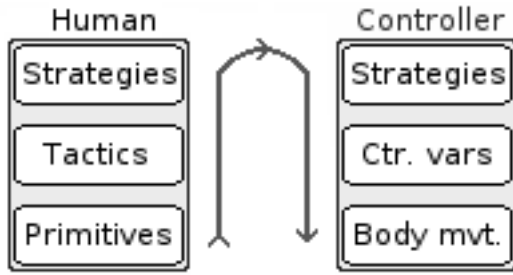


Figure 4. From primitives to body movements

the induced strategies into control variables (“Ctr. vars” in fig. 4) and robotic body movements (“Body mvt”).

## 6 A formal definition of “strategies”

A situation variable is a descriptor of perceptions of the environment external to the controller<sup>7</sup>. Each of the  $M$  situation variables has a finite and known number of possible values.

A control variable is a descriptor of robot action. Each of the  $N$  control variables has a finite and known number of possible values.

Formally, a robotic strategy is:

- A finite set of external situation states,  $E$ . Each situation state of  $E$  is expressed by a vector of  $M$  situation variables values:  $(e_1, \dots, e_M)$ .
- A finite set of internal action states,  $I$ . Each action state of  $I$  is expressed by a vector of  $N$  control variables values:  $(i_1, \dots, i_N)$ .
- An action transition matrix mapping all possible situation states to all possible action states. The values contained in this matrix are the probabilities of the robot enacting the behaviour described by an action state given a situation. We call it  $\Lambda_A = a_{ij}$ .
- An action duration mean transition matrix mapping all possible situation states to all possible action states. The values contained in this matrix are the means, should the robot enact the behaviour described by an action state, of the duration of all control variables of this behaviour. We call it  $\Lambda_{AD} = a_{d-ij}$ .
- An action duration standard deviation transition matrix mapping all possible situation states to all possible action states. The values contained in this matrix are the standard deviations, should the robot enact the behaviour described by an action state, of the duration of all control variables of this behaviour. We call it  $\Lambda_{ASD} = a_{sd-ij}$ .

Whenever the situation state of the robot changes, the robot goes into a certain action state chosen randomly according to the probabilities of  $\Lambda_A$ . It draws durations, in independent draws, for all the control variables values according to the Gaussian probability distributions defined by  $\Lambda_{AD}$  and  $\Lambda_{ASD}$  and starts a countdown to implement these durations.

When the situation state of the robot does not change but one of the control variable values reaches the end of its randomly assigned number of time steps, the robot goes into another action state chosen randomly, according to the probabilities of  $\Lambda_A$ , among all action states which have the same values for all the other control variables and a different value for the control variable which is due for a change. It draws a duration for this new control variable value according to the probability distributions defined by  $\Lambda_{AD}$  and  $\Lambda_{ASD}$ .

<sup>7</sup> Not necessarily “external to the robot”, the input from a sensor describing the state of the internal battery would be a situation variable value.

A discussion of this model and of how it compares to Hidden Markov Models can be found in [12], and there we explain that brute force learning of such a model would be intractable, not least of all because it would require enormous log files. We were not the first to say that such transition matrices are intractable, it has been known since the eighties, this is why we use several steps of generalisation and constrain our matrix. We cannot take one giant step from perception to cognition but we can climb this hill by taking several consecutive small steps, each time building upon the result of the previous generalisation step, and looking ahead to provide guidance to our learning algorithms.

## 7 Changes of strategies

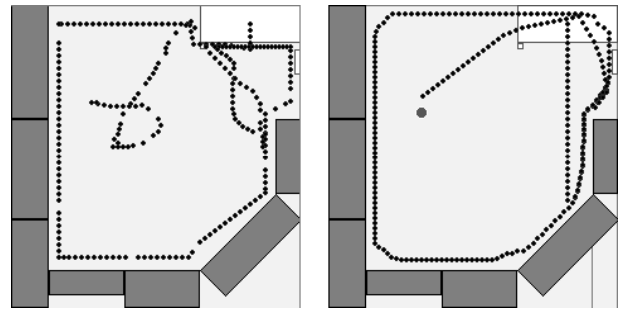


Figure 5. Track of G7.2’s run (left) and of our robot implementing her strategy (right)

How do humans recognise failure? In [12], we document an example of a volunteer, G7.2, changing her search strategy after coming across the same unique object for the second time: she apparently realised she was going round in a circle (fig. 5). Testing for possible failures can be part of a global search strategy, again not one which could be learnt in one step from reasonably-sized logs but one which requires a multi-levels control architecture and step-by-step learning. G7.2’s behaviour was learnt because from the observables “right hand  $(X, Y)$  coordinates”, “left hand  $(X, Y)$  coordinates” and “unique object number  $N$   $(X, Y)$  coordinates” the primitives “RHU” and “LHU”, right or left hand in contact with a unique object, were built<sup>8</sup>. During our statistical analysis (documented in [11]) the connection between coming in contact for the second time with a unique object and the “going off into empty space” behaviour, primitive of the “moving tactic”, became apparent. Failure recognition can lead to a change of strategy, which consecutive enaction can be observed.

In our formalism, switches from one action state to another can be triggered both externally, by a change of situation state, and internally, by the reaching the end of some control variable value life span. When the log file shows such a switch happening independently of these two conditions, it corresponds to a change of strategy. Changes of strategy are defined by a subset of situation states, either of which triggers the change, by a consecutive sequence of situation states belonging to this subset, or by a time limit assigned to each consecutive strategy (when they are programmed by hand).

<sup>8</sup> The observables are outside the controller, they are the raw data observed on the videos. The controller has no access to  $(X, Y)$  coordinates, it only has access to the knowledge that the right or the left hand is in contact with a unique object and encounters with a unique object are numbered.

We are interested in how robots can learn problem-solving strategies by human imitation. We are not interested in the strategies themselves. We deliberately make no selection related to the efficiency of the strategies our robot imitates. Our system observes subjects, analyses and implements not only winning strategies but all observable strategies.

Human beings have several “general purpose” problem solving strategies at their disposal [1]:

- Means - end analysis
- Working forwards / working backwards
- Induction
- Planning
- Trial and error
- Divide and conquer
- Etc...

We are interested in all of them, not only in winning strategies, because even if a given strategy, for example trial and error, may not be very efficient in our context, it is efficient in other contexts and the fact that it can be learnt by imitation is in itself important. And failures are important. The best-known example is probably the error backpropagation algorithm [8]. It is applied by hypothesis generate-and-test learning methods of all kinds. Learnt and/or evolved robot controllers with predictive capabilities test their predictions, sometimes at each time step, sometimes later on, sometimes only at the end of the run. Whole control architectures can be made to evolve when they fail some test.

The tests which failure drives learning and/or evolution range from the very specific hypothesis “at the next time step sensor 18 will have a value of 12” to the global “the goal will be reached”. In the mazes experiments, the volunteers resort to the global test, but also make up their own testable sub-goals as they go along. For example, a sub-goal can be to ascertain that the treasure is not on a given table before moving on to the next table. The failure of the corresponding test leads to backtracking and exploring the table again.

## 8 The H-CogAff architecture, the global alarm system and bumps

One of the more promising approaches to understanding what constitutes a mind seems to be modular multi-layered approaches such as H-CogAff.

Modularity enables divide-and-conquer problem-solving. So H-CogAff is implementable, albeit for the moment only in a very limited and schematic way, and can thus be tested.

Our tests showed us that because of its multi layers H-CogAff supports learning human problem-solving strategies by imitation, a human capability. This multi-layers characteristic enables step-by-step learning, solving an otherwise intractable problem.

The next set of tests, which this paper is about, showed us that adding a global alarm system (GAS) improved the performance of our simulated robot. The GAS does not only “make sense”, we proved that it actually works as predicted in our limited implementation of the H-CogAff model.

In the rest of this section we discuss parts of the H-CogAff, and in particular the GAS, comparing the model with our implementation of it. Testability being important for validating this model, we also discuss some possible observable effects that the GAS could have upon human beings. Finally we explain how we decided to test it and justify our decision by presenting the available data.

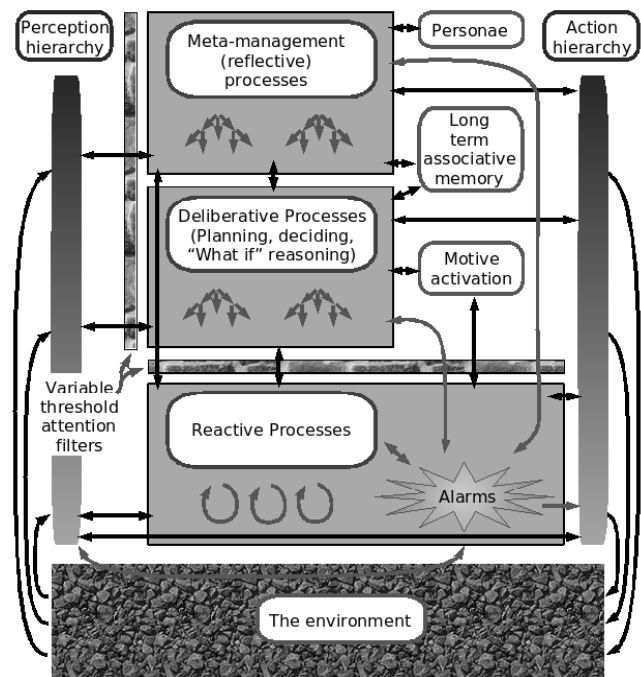


Figure 6. One version of H-CogAff

Fig. 6 is a more detailed representation of one version of the H-CogAff [21] architecture than the simplified one we showed in fig. 3. It has many similarities with robot control architectures (see for example [25] and [16] but many more exist which are just as interesting). In our previous model we implemented the environment, input, output and the three layers. Our statistical inference could be seen as corresponding to the long term associative memory. We absolutely do not claim to have implemented modules which even approach the complexity of the corresponding processes in the human brain. Our current model was built focusing upon the detection of changes of strategy, and H-CogAff was our inspiration.

According to H-CogAff, the human mind operates on three levels, each with its own functionality in terms of processing information. The lowest module senses the outside world, taking in stimuli from its infinite variety of sources, senses the inner state of the body and takes care of muscular control. It holds a summary of its perceptions and actions, written in a higher-level language, at the disposal of the module above. It has both a short term and a long term memory. The short-term memory is illustrated by play-back capacities: if we move our hand without being aware of it, maybe because we are shifting position, and put our hand down on something squishy, this both triggers the alarm system which draws our attention to the fact and offers us a playback of the movement we just did. Even if we automatically withdrew our hand after the event this playback enables us to know where it had been. The long term memory can store, probably in the very shape of the processes taking place in this module, the result of learning propagated down from higher modules. In natural language processing this lower module learns to do its equivalent of speech-to-text and text-to-speech. When we learn to drive we have to think about shifting gears but as we become more experienced this whole sequence of finely tuned muscular movements drops below the attention threshold. We drive on, secure in the knowledge that if something went wrong, if the gear level jammed, our lower module would immediately report failure via the alarm system. Without the learning abilities of the lower module even very day-to-day actions

such as brushing our teeth would become overwhelmingly complex.

In our system, learning was also required to define for example the type of trajectories the volunteers followed when walking across empty space. We could set the “class” variable<sup>9</sup> to 2 when the volunteer was walking approximately straight, to 1 when he or she was walking along an approximate curve and to 0 when he or she was approximately U-turning. Then the inductive algorithm<sup>10</sup> combined the values of differences in position and direction between current and shortly passed time steps and found good descriptions of these primitives.

The global alarm system is a mechanism enabling the lower-level processes, which for most of the time take place below the attention threshold, to suddenly draw attention to themselves when something goes wrong. A person can walk without thinking about it but if this person stumbles he or she will immediately pay attention to his or her feet and their immediate surroundings. The power of the global alarm system is such that it can “take over” the whole system, switching action control from the upper levels down to its own. If it detects imminent physical danger it can trigger a release of adrenaline, and people having experienced this report that afterwards they need a noticeable lapse of time before they can get their higher-level thought processes back to normal. The global alarm system can also send messages straight to the emotion centre (not a specific module in H-CogAff), for example when we are having a conversation, and paying attention to what is being said, it can suddenly report “the person you are talking to is very attractive”, with potentially disruptive consequences as far as the conversation is concerned. So the global alarm system can trigger a change of strategy, in this case the strategy changes because the goal changes, from “how can I convince this person that my point of view is better?” it becomes “how can I convince this person to have dinner with me?”. So an imitation system whose purpose would be to emulate this subject’s behaviour would have to learn two distinct strategies, the one which took place before the alarm and the one which took place after it. Detecting changes of strategy, in our case abrupt variations in the primitives values distributions, cannot be used to teach a robot’s global alarm system the set of circumstances in which it should send out an alarm until we have a way of differentiating the changes triggered by the alarm system from the changes triggered by some other process. But, given a set of alarm signals and the time step at which they occurred, we can test whether it makes sense to consider that the strategies taking place before and after the alarm are different, and this is what we did.

When the videos of the psychological experiments were converted to log files, variables were recorded which were later discarded both by our attribute construction algorithm and by us, for example the direction in which the head was facing. The purpose of the algorithm at that point was to construct primitives from observables and it had no look-ahead capacities which would have enabled it to estimate the usefulness of the observables upon the detection of changes of strategies: in our step-by-step approach the inductive algorithm was looking at the next module up and not right to the top.

One of the discarded observables was the binary “bump” variable. The “bump” observable took a value of 1 when the volunteer bumped into an obstacle or a wall and it’s value was 0 otherwise<sup>11</sup>. Among our

50 observable variables, “bump” was the one which seemed to have the highest likelihood of triggering the global alarm system of our volunteers. So we decided to set “bump” as a situation state which forced a change of strategy.

## 9 The variables which values control robot actions

These variables, called “control variables”, express by their values what the robot does. Our 6 control variables are the basic bricks from which the human-like strategies were implemented. We had 36 primitives built from 50 observables, and 4 tactics. We explain why there are now 6 control variables. The robot controller programming language that we implemented gives the actions which should be performed in all possible situations. One value of a control variable corresponds to one action of one part of the robot’s body. Our programming language expresses, in probabilistic terms, when a control variable value should change and to which new value it should change.

When we reached this stage we knew how to express the strategies in terms of tactics and the tactics in terms of primitives and observables. So we sorted our task-relevant actions according to body parts and possible contexts. The legs are considered a single body part. This was also a way of eliminating the conflicts which would certainly have arisen, in a probabilistic context, if we had attempted to implement the tactics independently one from the other.

The control variable are in the middle level, the lower level of the controller, “body movements”, does not need to know whether the robot should move forwards because it is part of its search tactic or because it is part of its obstacle following tactic.

### 9.1 EXPLORE\_OB: Object exploration behaviour

Exploring an object means in this context feeling its surface with one or two hands. As the robot goes along an obstacle, the hand on the side towards the obstacle can either follow the edge of this obstacle, sweep the obstacle, or stay by the body side. This control variable defines what the hands are doing relatively to an obstacle when the robot is near an obstacle. Sweeping a wall, apparently in search of top shelves, was rarely done by the volunteers and was in contradiction with the information the volunteers had been given, namely that the treasure would be on the ground or on a table but not on the ground under a table. So we did not implement this behaviour.

EXPLORE\_OB has five possible values:

1. None: The hands stay by the body side or searches the empty space in the obstacle detection behaviour.
2. 1HnE: One hand not efficient. One hand follows the near side of the object, only covering a small percentage of its surface as the robot goes along. The other one is a the body side or exploring the empty space.
3. 2HnE: Two hands not efficient. Both hands act as described above. They are following one another along the near edge of the obstacle the robot is following.
4. 1HE: One hand efficient. One hand sweeps the whole surface or nearly the whole surface as the robot goes along.
5. 2HE: Two hands, at least one acting efficiently.

### 9.2 EXPLORE\_GR: Ground exploration behaviour

There are three ways the robot can explore the ground. They can occur near an obstacle or in empty space, and can coincide with strictly

<sup>9</sup> In machine learning, the “class” variable is an exogenous variable which value specifies to which category the example belongs

<sup>10</sup> We tried a few and C4.5 [17] and [18] turned out to be the most efficient for our purpose

<sup>11</sup> We had many descriptors for the “obstacle following” behaviour but none in any of the higher modules which took into account whether the volunteer had come into contact with the obstacle gently or whether he or she had bumped into it.

positive values of the obstacle exploration variable, though when the ground exploration variable is equal to 4 the obstacle exploration variable can only be equal to 1 or 2.

EXPLORE\_GR has four possible values:

1. None: No exploration of the ground
2. Walk: The simplest, and least efficient ground exploration, is walking.
3. Sweep: The robot can “sweep” the ground, emulating the behaviour which consisted for a person of swinging his/her legs at each step to cover more ground than by walking.
4. Bend: The robot can emulate the behaviour of a person bending down or squatting and exploring the ground with his/her arms. This is the most efficient ground exploration technique.

### 9.3 EMPTY\_SPACE: Empty space behaviour

The volunteers in the mazes, when they ventured into empty space, strongly tended to either go approximately straight or to U-turn along a half circle of narrow radius. This control variable describes the trajectory in empty space.

EMPTY\_SPACE has four possible values:

1. Not\_Walk: Not walking across empty space.
2. Straight: Walking approximately straight across an empty space.
3. Curve: Walking along a approximate curve of large radius.
4. Turn: Walking along a approximate curve of narrow radius.

### 9.4 GO\_OFF: Going off into empty space behaviour

This variable controls what happens when the robot is near an obstacle in terms of will it stay near it or will it go away into empty space. Going from one obstacle to another, for example when following a row of tables, is not going off into empty space.

GO\_OFF has two possible values:

1. No: The robot does not go away from the obstacle
2. Yes: The robot chooses a random open direction and takes a step in that direction. It can go straight away from the obstacle or nearly tangentially to it, getting only marginally farther from it during the first steps.

### 9.5 FOLLOW\_OB: Obstacle following behaviour

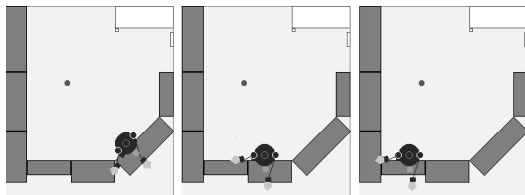


**Figure 7.** A time sequence of screenshots of a simulated robot enacting EXPLORE\_OB = 2HE and FOLLOW\_OB = Serious

The “Careless” vs. “Serious” qualifications for the obstacle following variable have to be understood as applying specifically to the obstacle following behaviour. Some people in the maze were seen to be sometimes following an obstacle, or a row of obstacles, quickly and efficiently. This could happen when they had just met a known

unique object and were backtracking away from it. They were then only interested in the obstacle as a guideline. Fig. 7 illustrates an obstacle following behaviour and an object exploration behaviour in the G7 maze. These screenshots do not illustrate consecutive time steps because our time steps are 0.25 second apart and the difference not so easy to see.

FOLLOW\_OB has three possible values:

1. Not\_Follow
2. Careless
3. Serious

When an obstacle is encountered, if the going off variable has a value of “No” and the obstacle following variable has a value of “Not\_Follow”, the robot will just stay where it is.

### 9.6 OB\_DETECTION: Obstacle detection behaviour



**Figure 8.** A volunteer whose actions would be expressed for the robot as EXPLORE\_OB = 1HnE and OB\_DETECTION = 1\_Hand

The blindfolded volunteers in the mazes often slowly waved their hands in the air, with their arms held horizontally or with their hands held approximately at the height of the last obstacle encountered. One example of this behaviour is shown in fig. 8. This behaviour is what the obstacle detection variable describes.

OB\_DETECTION has four possible values:

1. None
2. 1.Hand: When the person is following an obstacle and the obstacle exploration variable indicates that this obstacle is only being explored with one hand, it is still possible for the robot to do obstacle detection with the other hand.
3. 2.Hands: This can only happen when the person goes off or is in empty space.
4. Bump: “Bump” was implemented for completeness as a robot control variable possible option: we wanted all the observables to be reproducible, but the result of this control variable being set to “bump” was simply to make the robot rush straight into the first obstacle that happened to be in front of it and it was never included in any of our automatically generated robot control programs.

### 9.7 Implementation of human-like behaviour

A human-like behaviour is one which does not deviate by more than 3 standard deviations from the average of all recorded human behaviours (as shown in [11]). The control variables values are set according to average duration and not average value. This is illustrated by fig. 9: If the binary variable represents “Moving forwards”, the top part of fig. 9 would be a slow walk, the discretisation in time steps and integer distance units filling the log with alternating values 1 and

0. The lower part would be a fast walk followed by a pause. The average value of this variable is obviously the same top and bottom, one half, so these two very different behaviours can only be distinguished by the average durations of consecutive series of values.

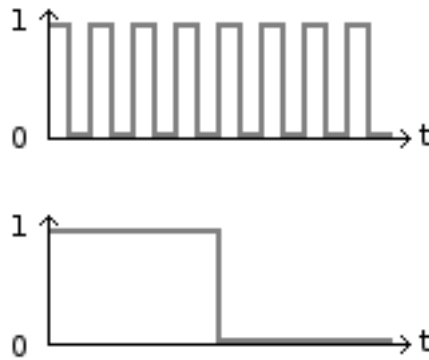


**Figure 9.** Same average, different behaviour

Learning was necessary during the bottom-up analysis, but, once the model was built, finding the average durations and their standard deviations to implement a given strategy only requires counting, given a specific context, the various corresponding occurrences in the log file. For example in the empty space context a volunteer could walk straight ahead for 12 time steps on average with a standard deviation of 3 and along a curving trajectory for 10 time steps on average with a standard deviation of 5. The values implementing this behaviour were randomly drawn with the corresponding probability distribution.

Without attempting to implement changes of strategies within a run of a volunteer in a maze the most efficient program was the one which considered all the 10 strategies together, independently of the duration of the run [12].

To implement our very limited version of the H-CogAff global alarm system, we set our program to divide the log files into several segments according to the bumps. As can be seen in table 1 the runs of 4 volunteers were not segmented because these persons never bumped into an obstacle. So these runs are considered to be driven by a single strategy all along. Twice, two bumps occurred shortly one after the other so the number of values between them was insufficient for a significant description of a strategy and the second bump was ignored. We ended up with 27 strategies which we combined.

**Table 1.** Durations and numbers of bumps

|       |        |        |        |       |       |
|-------|--------|--------|--------|-------|-------|
| Run   | G1.1   | G1.2   | G1.3   | G3.1  | G3.2  |
| Dur.  | 1'50"  | 31'48" | 12'53" | 1'33" | 1'44" |
| Bumps | 0      | 4      | 4      | 0     | 0     |
| Run   | G4.1   | G4.2   | G4.3   | G7.1  | G7.2  |
| Dur.  | 18'46" | 18'37" | 18'29" | 4'54" | 4'43" |
| Bumps | 6      | 1      | 2      | 2     | 0     |

Next we describe how we combined strategies in this case. If all the combined strategies are given equal importance, the length of the corresponding run notwithstanding, the result of dividing a run according to bump occurrences increases the relative importance of the strategies of the volunteers who bumped into obstacles more than others. In this case any improvement could be shrugged off with considerations such as a more active or even a more daring attitude is

beneficent to search efficiency as well as being detrimental to personal safety. So on one hand we combined all strategies occurring during one minute after a bump together into the “after bump” strategy, and on the other hand we combined the strategies taken from the rest of the runs into the “no bump” strategy. The one minute limit was chosen arbitrarily. When the robot bumped into an obstacle it abruptly switched to the “after bump” strategy, all probabilistic values being redrawn at once. One minute later the robot started implementing the “no bump” strategy but this time the change occurred more gradually, each variable finishing the life span assigned to its current value before drawing a new value according to the “no bump” strategy.

## 10 Experimental results

We consider that the fact that some human problem-solving strategies are learnable by a robot is more important than the actual strategies being learnt. It must also be remembered that the experimental settings were strongly constrained and that using bumps as the sole trigger of our “global alarm system” does not do justice to this complex system.

We call the result of the program generated from all logs indiscriminately joined together the result of robot.1 and the result of the program generated from the two distinct sets “after bump” and “no bump” the result of robot.2.

### 10.1 Reasons for the lack of comparisons between robot and subject performance

The task of the volunteers was to find the treasure (or, in some cases, for one of them to find the treasure, after which they could all walk out). So their performance in terms of time to achievement and/or ground coverage depended too much upon the original location of the treasure for performance comparisons to be made. We only had the logs of ten subject runs. Each subject only went through a maze once. They did not all go through the same maze. Three subjects definitely did not spend all their time in the maze searching for the treasure. So such comparisons would in any case have had low statistical significance.

The logs of successful volunteers who found the treasure quickly, do not generate better controllers. In fact in one case it was quite the opposite: G1.1 used a systematic right-handed blind man walk: she followed the wall on her right, thoroughly exploring all objects she encountered on her right and ignoring anything on her left. She quickly found the treasure because it was located on the ground, near a radiator which was fixed to the wall. The corresponding controller makes the robot go round in circles following the outside perimeter of the maze and never explores the interior so it is not a good sweeper. G1.1 only covered about a tenth of the maze before finding the treasure. In terms of time to achievement she was fast, in terms of ground coverage she was very slow.

On average the subjects took 11'32" to find (or for one of them to find) the treasure. Robot.1 reached between 78% and 94% of all reachable squares at least once after 10 minutes and robot.2 reached between 80% and 94% of all reachable squares at least once after 10 minutes. Concluding that, given uniformly random treasure locations, the robots did better than the subjects would be, to say the least, premature.

Given these facts and lack of a lot more of human-related data we regretfully forwent subject-to-robot comparison measurements.

## 10.2 Robot results

In the following scale and speed correspond to the real settings.

The settings for the following were the four mazes from which one or several logs had been drawn, and six extra (invented) mazes used for testing purposes.

- The performances were not better in the “known” mazes than in the invented mazes, showing that the strategies had really been abstracted from their original settings.
- All tables had been explored after at most 11 minutes by robot\_1. After 10 minutes robot\_2 had explored all tables in all but one maze. That last maze had an isolated central table which robot\_2 often bypassed in its exploration of the empty space.
- On average, 83% of the tables had been explored after 3 minutes by robot\_1 and 86% by robot\_2.
- Dividing the ground in squares 20 pixels across<sup>12</sup>, which corresponds to the average “width” of a subject as seen on the videos, between 78% and 94% of all reachable squares (ground and tables) had been reached at least once after 10 minutes by robot\_1, and between 80% and 94% by robot\_2, the actual average values varying according to maze size and complexity. The improvement mostly happened in complex mazes with many tables.
- These percentages increase with run duration.

The difference between the efficiency of robot\_1 and of robot\_2 might seem slight, but as performance improves each percent point becomes harder to gain than the previous one. So the difference between 83% and 86% of the tables is more significant than would have been a difference between 63% and 66% of the tables.

## 11 Conclusion

Closing the loop: As our robots move about in the mazes, their observable actions can be recorded. Thus the teacher/learner/teacher loop, sometimes also referred to as the raw-data/trained-system/raw data (or raw-data/learned-strategies/raw-data) loop, is closed because new logs can be generated and learning can be achieved from these new logs. These new logs are neither a better nor a worse model than the originals, were the robot generating them to run for a sufficiently long time and in a sufficiently complex maze for the randomness to be overcome by statistical significance they would amount to the same information.

H-CogAff validation: H-CogAff describes the functions of the mind [22]. Our experiments only validate the H-CogAff model in a severely limited and constrained context. They also validate the global alarm system, with the same restrictions and more because a single variable was used to trigger strategy changes, and this variable was neither learnt nor even statistically determined but was chosen by empirical reasoning. We can however say that an instant reaction which resets every control variable upon a bump improves performance, and that this seems to indicate that simulating a global alarm system, hopefully in more advanced ways, in other robotic controllers could be a promising line of research.

## ACKNOWLEDGEMENTS

Our link between the global alarm system and emotions was first suggested by David A. Focil.

<sup>12</sup> We were given the plans of the mazes by the psychologists, but these plans had no scale. When we inquired about it we were told that the corresponding rooms still existed and could be measured. We declined and used pixels for our distance unit.

## REFERENCES

- [1] Jean-Francois Richard Aldo Zanga and Charles Tijus, ‘Implicit learning in rule induction and problem solving’, *Thinking and Reasoning*, **10** (1), 55–83, (2004).
- [2] A. Alissandrakis, C. L. Nehaniv, and K. Dautenhahn, ‘Learning to imitate corresponding actions across dissimilar embodiments’, *IEEE Transactions on Systems, Man, and Cybernetics*, **32**, 482–496, (2002).
- [3] A. Alissandrakis, C. L. Nehaniv, and K. Dautenhahn, ‘Correspondence mapping induced state and action metrics for robotic imitation’, *Special Issue on Robot Learning by Observation, Demonstration and Imitation*, (2006).
- [4] Billard and Hayes, ‘Drama, a connectionist architecture for control and learning in autonomous robots’, *Adaptive Behaviour*, **7**, 35–64, (1999).
- [5] A. Billard and R. Siegwart, ‘Robot learning from demonstration’, *Robotics and Autonomous Systems*, **47**, 65–67, (2004).
- [6] S. Calinon and A. Billard, *Imitation and Social Learning in robots, Humans and Animals: Behavioural, Social and Communicative Dimensions*, Cambridge University Press, 2007.
- [7] S. Calinon, F. Guenter, and A. Billard, ‘On learning, representing and generalising a task in a humanoid robot’, *IEEE Transactions on Systems, Man and Cybernetics, Part B, Special Issue on Robot Learning by Observation, Demonstration and Imitation*, **37** (2), (2007).
- [8] G. E. Hinton D. E. Rumelhart and R. J. Williams, ‘Learning internal representations by error propagation’, *David E. Rumelhart and James A. McClelland*, **1**, (1986).
- [9] Demiris and Hayes, *Imitation in animals and artifacts*, 2001.
- [10] R. Dillmann, ‘Teaching and learning of robot tasks via observation of human performance’, *Robotics and Autonomous Systems*, **47**, 109–116, (2004).
- [11] M. Felkin, ‘Learning by observation and induction the strategies of humans placed in a problem-solving context’, *PhD Thesis*, (2008).
- [12] M. Felkin and Y. Kodratoff, ‘High level imitation learning in a treasure hunting task’, *Proceedings of Plan, Activity, and Intent Recognition (PAIR) workshop, International Joint Conferences on Artificial Intelligence (IJCAI 2009)*, (2009).
- [13] Valentina Iemmi, ‘Les apports des nouvelles technologies a la psychologie clinique: Les robots comme compagnons therapeutiques’, (2005).
- [14] Heumer G. Jung B., Ben Amor H. and Weber M., ‘From motion capture to action capture: A review of imitation learning techniques and their application to vr-based character animation’, *Proceedings ACM VRST*, (2006).
- [15] Henri Kraus, ‘A formal theory of plan recognition and its implementation’, *Reasoning About Plans*, by J.F. Allen, H.A. Kautz, R.N. Pelavin, and J.D. Tenenbergs, 69–126, (1991).
- [16] J. P. Müller, ‘A conceptual model of agent interaction’, *Second International Working Conference on Cooperating Knowledge Based Systems (CKBS-94)*, 389–404, (1994).
- [17] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.
- [18] J. R. Quinlan, ‘Improved use of continuous attributes in c4.5’, *Journal of Artificial Intelligence Research*, **4**, 77–90, (1996).
- [19] S. Schaal, A. Ijspeert, and A. Billard, ‘Computational approaches to motor learning by imitation’, *Philosophical Transactions of the Royal Society of London: series B, Biological Sciences*, **358**, 537–547, (2003).
- [20] Storz J. Shon A. and Rao R., ‘Towards a real-time bayesian imitation system for a humanoid robot’, *Proceedings of the IEEE International Conference on robotics and Automation*, 2847–2852, (2007).
- [21] A. Sloman, ‘Varieties of affect and the cogaff architecture schema’, *Proceedings Symposium on Emotion, Cognition, and Affective Computing AISB’01 Convention*, 39–48, (2001).
- [22] Aaron Sloman, ‘Some requirements for human-like robots: Why the recent over-emphasis on embodiment has held up progress’, *Creating Brain-Like Intelligence, LNAI 5436*, 248–277, (2009).
- [23] C. Tijus, N. Bredeche, Y. Kodratoff, M. Felkin, C. Hartland, V. Besson, and E. Zietti, ‘Human heuristics for a team of mobile robots’, *Proceedings of the 5th International Conference on Research, Innovation and Vision for the Future (RIVF’07)*, (2007).
- [24] Mark P. Woodward and Robert J. Wood, ‘Using bayesian inference to learn high-level tasks from a human teacher’, *International Conference on Artificial Intelligence and Pattern Recognition (AIPR-09)*, (2009).
- [25] Victor Lesser Xiaojin Zhang and Tom Wagner, ‘A layered approach to complex negotiations’, **2**(2), 91–104, (2004).

# Grounding Symbols: Learning and Grounding 2D Shapes using Cell Assemblies that emerge from fLIF Neurons

Fawad Jamshed<sup>1</sup> and Christian Huyck<sup>1</sup>

**Abstract.** Symbols are the essence of a language and need to be grounded for effective information passing including interaction and perception. A language is grounded in a multimodal sensory motor experience with the physical world. The system described in this paper acts in a simple virtual world. The system uses multimodal interaction by using vision and language, to interact with the outer world. Two simulations are given in which the system learns to categorise five basic shapes. In the second simulation, the system labels these categories, grounding them. Both simulations performed work perfectly on all tests.

## 1 Introduction

Although many aspects of human cognition and language are still a mystery, guidance can be taken from human processing where possible [1]. Infants ground their language by interaction with the physical world, associating speech patterns with objects and actions in the world [2]. An infant initially learns categories through observing the world. Humans use a very complex communication system. It is widely accepted, that the human communication system is symbolic, learnt, compositional and multimodal. Understanding the underlying principles of the human communication system requires an understanding of the mechanisms with which the words' meanings are rooted in reality.

The system, whether it is human or artificial, must learn to categorize input into bins. So, when an agent senses an item, it must categorise it as, for instance, a dog. This allows the system to generalize its behaviour so that it can react similarly to different instances of that category. One explanation of this process of category learning is provided by Hebb [3]. A cell assembly, a reverberating neural circuit, represents the category or concept. The cell assembly is learned by modifying synaptic weights using a Hebbian learning rule, which states the synaptic weight between two co-firing neurons is increased. The categories are formed in response to environmental stimuli. Thus, the reverberating circuits of cell assemblies are formed by groups of features that tend to travel around together. That is, the cell assembly for dog is formed by a group of features that dogs tend to have. When a dog is presented to the system, the dog cell assembly becomes active. This cell assembly formation is a form of sensory motor toil [4]. New categories are learned

formed through real time, feedback-corrected, trial and error; the agent uses senses and motion to develop those categories.

These concepts are entirely subsymbolic. They are formed in direct response to the environment, and can be formed by animals that do not process symbols. However, humans do process symbols, and these symbols can be linked to the sub symbolic cell assemblies. This process of labelling is essential for symbol grounding.

A great deal of artificial intelligence concentrates on symbol processing, but no simple symbol processing system can be truly intelligent [5]. Instead, a host of implicit knowledge is needed, and that knowledge is stored in the underlying cell assemblies, and relations between cell assemblies. Thus, a truly intelligent artificial system must have grounded symbols [6]. Once a system has some grounded symbols, it can then use symbols to learn new symbols and new categories [7]. This is the benefit of language, a person can learn from language without actually having to experience the environment. An agent can learn where someone lives by being told their address. This symbolic theft, based on the categories that have been derived via sensory motor toil, makes the system much more powerful. According to symbolic theft, new symbols are formed by combining already grounded symbols [4]. Some basic categories must still be learned by sensory motor toil, and then by using symbolic theft, new symbols can be acquired.

Symbols are grounded by multimodal interaction with the world [8]. Multimodal communication can involve vision and language integration to ground mental concepts in the physical world [9]. This multimodal integration, between language and visual representation enables humans to acquire and use words in context and makes communication possible by establishing coherence between mental states and the physical world. Thus not only visual items need to be learned, they need to be labelled to ground them.

In this paper, a system that learns to categorise visual objects by sensory toil is presented. The system then labels these categories, providing the basic structure of symbol grounding. This system is implemented entirely in simulated neurons (see section 2.1). Cell assemblies (see section 2.2) emerge from the system's interaction with the environment. The learning of these visual items is described and evaluated in section 3. These items are then labelled; this process is described and evaluated in section 4. Conclusions are then drawn and future work discussed.

---

<sup>1</sup> School Of Engineering and Information Sciences- Middlesex University, The Burroughs –London, UK.  
Email: {f.jamshed, c.huyck}@mdx.ac.uk.

## 2 Theoretical Background and Previous Work

The system is based on a simple, but realistic, neural model. This model is the fatiguing Leaky Integrate and Fire neuron. In response to environmental stimuli, these simulated neurons can form cell assemblies. The precise form of the response is governed by particular Hebbian learning rules.

### 2.1 Fatigue Leaky Integrate and Fire Neurons

The fatiguing Leaky Integrate and Fire (fLIF) neuron is a relatively simple, though accurate, model of biological neurons. The fLIF neuron is an extension of the more widely used LIF model [10]. This is, in turn, an extension of the integrate and fire model [11].

The model used in this paper makes use of discrete cycles. A cursory description is provided and a fuller explanation can be found elsewhere [12]. Each neuron has some activation, which it receives from other neurons. If a neuron has enough activation at the end of a cycle, it will fire, spread activation to connected neurons, and lose all its energy. Neurons are connected to other neurons with unidirectional, weighted connections. A firing neuron passes the activation of the weight of the connection. If a neuron's activation is less than the threshold, it will not fire but some the retained activation will leak away for the next cycle. This is modelled by calculating the activation as describe in equation 1.

$$A(t) = A(t-1)/D + C. \quad (\text{Equation 1})$$

Where  $A(t)$  is the activation at time  $t$ . It is the sum of the activation at time  $t-1$ , reduced by a decay constant  $D$ , and  $C$ , the sum of incoming activation of all active neurons that are connected to a given neuron and fired at time  $t-1$ . The value of  $C$  is determined by multiplying the incoming activation from other firing neurons at time  $t-1$  with the associated weights of those links.

fLIF neurons also fatigue just like biological neurons. If a neuron fires regularly it becomes harder to fire. This is modelled by increasing the threshold of a neuron when a neuron fires, then reducing it when the neuron does not fire.

### 2.2 Cell Assemblies

A Cell Assembly (CA) is the neural basis of a symbol [3]. It is a subset of neurons that have high mutual synaptic strength enabling neurons in the assembly to persistently fire after external stimulation ceases. In the simulations discussed in this paper, a small subset of all the neurons represents a symbol. If many of the neurons in it are firing, the symbol is active.

Cell assemblies give a sound answer to the neural representation of two types of memory, long-term memory and short-term (or working) memory. The firing of many neurons in an assembly is the neural implementation of short-term memory; this high frequency and persistent firing makes the cell assembly active.

### 2.3 Hebbian Learning

The basic learning mechanism used in these experiments is the compensatory learning rule that has been described elsewhere [13]. The overall rule is derived from a correlatory rule that combines Hebbian and anti-Hebbian mechanisms. That is, the synaptic weight between two neurons increases when the neurons co-fire, and decreases when the pre-synaptic neuron fires and the post-synaptic neuron does not. Over time, the synaptic weight reflects the likelihood that the post-synaptic neuron fires when the pre-synaptic neuron does. This correlatory mechanism uses a modifier that forces that combined synaptic weight leaving a neuron toward a constant. This is done by making the increases smaller and the decreases larger when the total synaptic weight of a neuron is greater than constant; and making the increases larger and the decreases smaller when the total synaptic weight is less than constant. This rule makes synaptic weight a resource.

### 2.4 CABot

The simulations described in section 3 and section 4 are an extension of the first version of the Cell Assembly Robot (CABot1) [14]. The main aim of CABot is to develop an agent in simulated neurons, which can take natural language as input and interact with the environment without any external help. By interacting with the environment, it is hoped that it can learn the semantics of the environment sufficiently well to improve language processing.

For CABot1, a virtual 3D environment was established based on the Crystal Space games engine. Two agents were placed in the environment, the first controlled by a user, and the second was the CABot1 agent. All processing in CABot1 was done by a complex network of fLIF neurons, though it emitted symbolic commands to the Crystal Space stub.

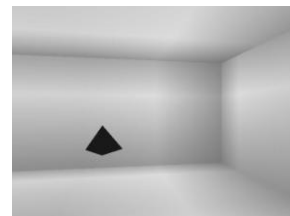


Figure1. Instances of pyramid

A complete description of CABot1 is beyond the scope of this paper but further information can be found elsewhere [15]. A total of 21 sub-networks are used to subdivide the tasks of vision, natural language parsing, planning, action and system control.

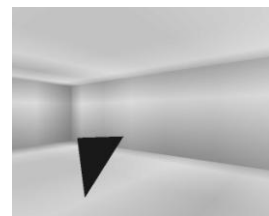


Figure2. Instances of stalactite

The important subnets for the purposes of this paper are the vision nets and the word nets. There are three vision subnets, a simulated retina, a primary visual (V1) cortex and a secondary visual cortex (V2). These systems were hard coded, so there was no learning. Visual input was in the form of a bitmap representation of a view of the game from the agent's perspective. In particular, the secondary visual cortex subnet was set to recognise *pyramids* and *stalactites*. If one of these was present in the game, a CA in V2 ignited. There were several position and size dependent CAs associated with both *pyramid* and *stalactite*. Figure 1 and figure 2 shows instance of *pyramid* and *stalactite* respectively.

Similarly, the parsing component had CAs for words. In the game, the user issues natural language commands to tell the agent what to do. There was a noun instance subnet to store semantic roles during parsing.

### 3 Learning Visual Items

Learning was introduced in to the system with the help of six visual sub nets. The five shapes used are: pyramid, stalactite, diamond, square and right angle triangle. Currently the vision system consists of six nets which are Input net, Retina net, V1 net, V1A net, V2 net and V4 net. Each of these six sub nets performs a unique function, takes input only from the prior net (the environment for Input), and passes input to the subsequent net.

The Input net displays the input from the environment whereas the Retina net is a series of OnOff and OffOn detector; these nets are unmodified from CABot1. The V1 net is position dependent and detects first order features of a solid shape in the picture. These features included including oriented edges and angles, and several new angle detectors were added from CABot1. The V1A net is a position independent model of the V1 net, merely showing the presence of the feature without location. The V2 net detects the second order features, a three way combination of edges and angles. For example, it detects a horizontal edge, a right angle, and a vertical edge are all simultaneously present. There were 13 such second order detectors. The V4 net identifies the shape of an object with the help of the second order features which are detected in the V2 net. The detailed working of the vision system is defined below.

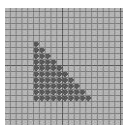


Figure 3a. Right Triangle

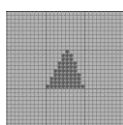


Figure 3b. Pyramid

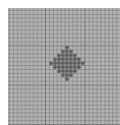


Figure 3c. Diamond

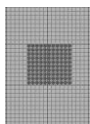


Figure 3d. Square

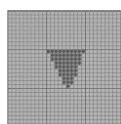


Figure 3e. Stalactite

Figure 3a, 3b, 3c, 3d, and 3e shows the instances of right angle triangle, pyramid, diamond, square and stalactite respectively.

The Input net gets the input from the system in the shape of bits and displays it on the screen. The input is usually in the shape of pictures but shapes can be hard coded in the system.

The Retina net is a biological plausible model of the OnOff and OffOn detectors that are found in biological systems; it gets the input from input net, and feeds its output to the V1 net. Three different types of OnOff and Off On detectors are used in the Retina net 3 by 3, 6 by 6 and 9 by 9 detectors.

V1 is position dependent and gets the input from the OnOff detectors and identifies the first order features e.g. edges and angles of a solid shape in the picture. The V1 net responds to different types of edges and angles presented. The connections from the V1 net were made position independent by introducing the V1A net and making random connections from each V1 CA to the corresponding V1A CA. The V1A net has direct connections from the V1 net only. The neurons of the V1A net have a low decay rate of 1.01 to promote even a small firing of neurons in V1.

The V1 net and the retina net used are modified versions of the V1net and the retina net of CABot1. More CAs are introduced in the V1 net including vertical edge CAs, and four types of right angle CA. Whereas the V2 net, the V4 net and the VT net (described below) were introduced for this experiment.

The V2 net gets input from the V1A net, which is the position independent version of the V1 Net. When a three or four edged shape is presented, each CA of the V2 net gets three inputs from three CAs of the V1A net. CAs of the V2 net are only ignited when all of three CAs in the V1A net is active when a shape is presented. The V2 net output is used as an input to the V4 net where final shape is determined.

The V4 net is the final part of the vision system where all the shapes are discriminated. The V4 net and the V2 net are fully connected which means each of the CAs in V2 net is connected with all CAs of V4 and vice versa. Learning is carried out between V2 and V4 nets.

The same topology was used within V1A net, V2 net and V4 net. In this topology, twenty percent of the neurons used were inhibitory neurons while eighty percent of the neurons used were excitatory. In the case of the V4 net, there are inhibitory intra-CA connections to promote winner takes all situations so that only one CA will eventually ignite.

In order to prevent a CA from winning on two different shapes, a slow learning approach is used. Due to the low learning rate, the weights do not grow rapidly. Hence the V3 net CAs does not commit to the V4 net rapidly, but over a longer period, reducing the likelihood that they will form incorrect associations.

### 3.1 Experiment

During the simulations, all five shapes are learned. Instances of those shapes are presented as visual input. The input neurons are clamped on for 250 cycles. Activation passes through the retina, V1, V1A and V2 subnets. The activation then passes to the V4 net; as the weights from V2 to V4 are initially low, no V4 CA ignites initially. However, each of these nets slowly accumulates activation, and eventually one ignites, and persists. As it inhibits the other CAs, it is the only one that persists. The winner takes all strategy prevents one shape being committed to more than one CA of the V4 net. During this phase of persistent firing, the weights from V2 and back to V2 are raised via Hebbian learning.

An instance of each of the five shapes is presented to train the network. It is then tested by presentation of other instances of the shapes. Table 1 shows the result of a successful test when an

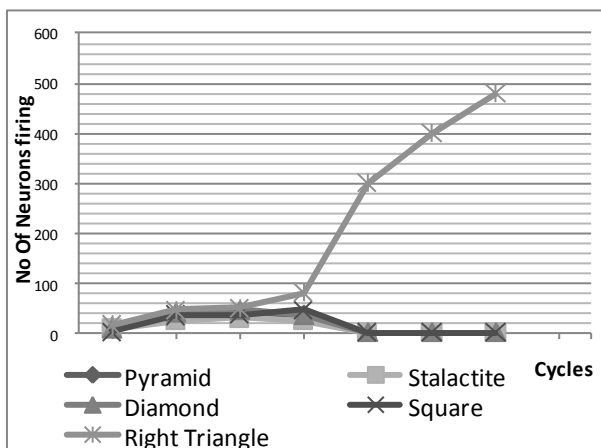
instance of visual pyramid and of visual diamond are presented; the number of neurons firing in the corresponding CAs of V4 show that specific CAs are committed to pyramid and diamond respectively, whereas the other CAs of the V4 net do not respond. The correct ignition of the CAs in the V4 net shows the system has learned these shapes.

**Table 1.** No of neurons firing in the V4 net during testing phase

| Neurons Firing            | Pyramid | Stalactite | Square | Diamond | Triangle |
|---------------------------|---------|------------|--------|---------|----------|
| When pyramid is presented | 490     | 0          | 0      | 0       | 0        |
| When Diamond is presented | 0       | 0          | 0      | 479     | 0        |

The simulation is termed successful when all of the five CAs of the V4 net are committed to the five different distinctive shapes, whereas the quality of the success is determined by how well these five shapes get learned CAs to respond when different shapes are presented.

**Graph 1.** Example of winner takes all in this case its triangle



Graph 1 shows a typical example of winner takes all strategy (in this case its triangle). Initially all the CA compete with each other but slowly one CA gets promoted which inhibits other CA neurons making them more harder to fire.

The training runs for 1250 cycles where each of the five shapes was presented for 250 cycles. The system is reset at 250 cycle intervals with all neural activation and fatigue reset to 0. The connections between the V2 net CAs and the corresponding V4 net CA are adjusted and learned using Hebbian learning. The learning is bi-directional with weights on connections from the V2 net to the V4 net and weights from the V4 net to V2 net being learned at the same time.

The test runs for 2500 steps. During the testing part of the simulation, shapes were presented in a random order. After presenting the complete set of five shapes in 1250 cycles,

another set of shapes was presented, randomly again, for another 1250 cycles.

## 3.2 Results

The test is fully automated and runs without human intervention. The test was conducted 28 times. The result is calculated using the correct numbers of corresponding CAs of the V4 net being fired when a shape is presented to the system.

The success rate among three shapes diamond, square and rectangle is one hundred percent. Each of the shapes i.e. diamond, square and rectangle, gets committed to a different CA each time they are presented and during testing each of the committed CAs responds correctly to different shapes presented each time. Due to the very similar features, pyramid and stalactite shapes do get committed wrongly sometimes and the CA which is being committed to one of the shapes presented first also sometimes responds to the second shape presented afterwards. The success rate among the shapes of pyramid and stalactite is 75 percent.

## 4 Labelling Items

Labelling is a simple but crucial part of the symbol grounding problem and associates the correct labels with semantic categories [16]. This simulation shows how learned semantic categories for shapes get appropriate labels attached to them. These associations are then tested. The use of semantic CAs and label CAs is consistent with linguistic research where a word has a semantic pole and a phonological pole [17].

### 4.1 Experiment

The system is presented with different pictures of instances of shapes and after detecting the appropriate first and second order features, the correct shapes are learned (as described in section 3). These learned shapes are then associated with a label by presenting an instance of shapes and its label simultaneously. The connections are adjusted and learned using Hebbian learning.

Learning is bi directional so the label can be used to retrieve the shape, and the shape can be used to retrieve the label. Each shape and label is represented by a CA of 500 neurons in the V4net and the label net respectively. Every 5th neuron in each of shape and label CAs has 3 random connections with the CAs of each label in the label net and with the CAs of each shape in the V4net. The values of other parameters used in the experiment are as shown in the table 2.

**Table 2.** Values of the parameters used in the experiment

| Parameter             | Value |
|-----------------------|-------|
| Decay rate            | 1.45  |
| Learning rate         | 0.2   |
| Fatigue Rate          | 1.2   |
| Fatigue Recovery rate | 1.4   |
| Activation threshold  | 4.2   |

The correlatory learning rule is used to learn the connections. According to the correlatory learning rule, when pre-synaptic

and post-synaptic neurons fire, the weight is increased, and when the pre-synaptic neuron fires and the post-synaptic neuron does not fire the weight is decreased. Initial weights of connections between the shapes and the labels are set low (0.01) so the weight can be adjusted to the appropriate level.

During testing, the semantic input to the system is turned off to check if the right semantic CA is ignited when a label is presented. Similarly to test the connections from the semantics to labels, input to the label net is turned off to see if the correct label ignites.

During learning, each instance of shape and label was presented for 250 cycles. Each epoch of learning and testing for all the shapes and the labels lasted for 1250 cycles. After 4 epochs of learning the system is tested for 4 epochs. In the first 2 epochs of testing, the semantic input to the system is turned off to test the label to semantic connections. In the next 2 epochs of testing, the semantic input to the system is turned on but the input to the label net is turned off to test semantic to label connections.

**Table 3a.** Visual pyramid is presented when Input to the label net is off.

|                             | Pyramid | Stalactite | Square | Diamond | Triangle | Label Pyramid | Label Stalactite | Label square | Label Diamond | Label Triangle |
|-----------------------------|---------|------------|--------|---------|----------|---------------|------------------|--------------|---------------|----------------|
| <b>No Of Neurons Firing</b> | 480     | 0          | 0      | 0       | 0        | 459           | 0                | 0            | 0             | 0              |

**Table 3b.** Label stalactite is presented when Input to the V4 net is cut off.

|                             | Pyramid | Stalactite | Square | Diamond | Triangle | Label Pyramid | Label Stalactite | Label square | Label Diamond | Label Triangle |
|-----------------------------|---------|------------|--------|---------|----------|---------------|------------------|--------------|---------------|----------------|
| <b>No Of Neurons Firing</b> | 0       | 465        | 0      | 0       | 0        | 0             | 472              | 0            | 0             | 0              |

Table 3a shows the result of a run when the input to the Label net is switched off and an instance of visual pyramid is presented to the system. When an input is presented in V2, a CA in the V4 net is ignited, which, in turn, ignites the correct associated Label CA even though it has no external input. Table 3b shows the results of a run when the visual input to the system is turned off and input is applied to the Label net. The CAs of the Label net also ignites the corresponding CAs of the V4 net even though there is no visual input to the V4 net. These tables clearly show the appropriate connections are being learned.

## 4.2 Results

The results were calculated by counting the number of neurons fired at 249<sup>th</sup> cycle of presenting a shape or label. During learning and testing different instances of the shapes were used. On all 10 tests that were carried out, all the right CAs ignited

when the system is presented with the corresponding CAs. This correct ignition of the CAs shows the correct association between shape and label CAs is learned appropriately.

## 5 Conclusions and Future Work

As has been pointed out, to some extent the Symbol Grounding Problem has already been solved [18]. That is, artificial systems have learned categories and labelled them. Of course, the problem still remains unsolved in that no artificial system currently grounds symbols nearly as effectively as a typical human does.

This paper describes an early step toward symbol grounding. Like people, this system processes using neurons. Also like people, this system learns new visual categories. Finally, like people, this system labels those categories.

As compared to other models (for example [19], [20], [21], [22]), our model uses biologically plausible neurons, and the system is more generic in nature (as it is not only bound to spoken words). The model described in this paper has the flexibility to be used with incremental learning and can use environmental feedback to create or readjust what is already learned [23].

This paper has shown progress on the symbol grounding problem. The symbol grounding problem is clearly a problem that humans must solve, but little work in the study of natural cognition centres on this issue. This is probably because this is so effortless for humans. Humans learn new categories and instances constantly throughout life. They do this implicitly and explicitly; those memories may be labelled or not; and the labelling is also usually effortless. However, as greater understanding of natural cognition evolves, these complexities will be better understood.

The promising results of these experiment shows that the technique and model used for these experiments can be used effectively to solve other aspects of the symbol grounding problem. Other aspects of that need to be addressed in order to solve the SGP are functional symbol grounding, symbolic theft, hooking, alignment and environmental feedback. Functional symbol grounding involves grounding a symbol in a given context. Symbolic theft involves grounding new categories from splitting or merging already grounded symbols. For example by combining stripes with horse, a new category, zebra, can be created and vice versa. Hooking is when a category can have more than one representation; for example a chair looks very different from the front and back. An intelligent agent should be able to ground all representations such that object is categorised regardless of the angle of input. Hooking also refers to synonymy [24] with multiple labels applying to a single object or category. Alignment involves coping with an unfamiliar symbol by using the information from already grounded symbols. For example if an agent has not seen an object before, using its already grounded symbols, the agent can infer the meaning of the new symbol. Environmental feedback means an agent should be able to learn or readjust already grounded symbols.

Though, this list is not exhaustive, addressing these issues will yield an agent that can much more effectively ground symbols. Such an agent will be able to make its own decisions without little help from the outside world.

## References

- [1] Clark, A. (1997). "Being There: Putting Brain, Body, and World Together Again" Cambridge, MA: MIT press.
- [2] Bloom, P. (2000). "How Children Learn the meaning of Words" Cambridge, MA: MIT Press.
- [3] Hebb, D. (1949). *The Organization of Behavior*. New York: Wiley.
- [4] Cangelosi, A., Greco, A., & Harnad, a. S. (2000). "From Robotic Toil to Symbolic Theft: Grounding Transfer from Entry-Level to Higher-Level Categories" *Connection Science* , 143-162.
- [5] Searle, J. (1980). *Minds, Brains and Programs*. Behavioral and Brain Sciences , 417-424.
- [6] Harnad, S. (1990). "The Symbol Grounding Problem". *Physica D* , 335-346.
- [7] Massé, A. B., Chicoisne, G., Gargouri, Y., Harnad, S., Picard, O., and Marcotte, O. (2008). "How is meaning grounded in dictionary definitions?" In *Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms For Natural Language Processing*. ACL Workshops. Association for Computational Linguistics, Morristown, NJ, 17-24.
- [8] Dillenbourg, P., Traum, D., and Schneider, D. (1996). "Grounding in Multi-modal Task-Oriented Collaboration" In *proceedings of the Euro AI and Education Conference*.
- [9] Pastra, K. (2004). "Viewing Vision-Language Integration as a Double-Grounding Case" In *Proceedings of the American Association of Artificial Intelligence (AAAI) Fall Symposium on "Achieving Human-Level Intelligence through Integrated Systems and Research"*, pp. 62-67, Washington D.C. USA.
- [10] Amit, D. (1989). "Modelling Brain Function: the world of attractor networks". Cambridge University Press.
- [11] McCulloch, W., & Pitts, W. (1943). "A Logical Calculus of Ideas Immanent in Nervous Activity". *Bulletin of Mathematical Biophysics* , 115-133.
- [12] Huyck, C. (2007). "Creating Hierarchical Categories Using Cell Assemblies." *Connection Science* , 1-24.
- [13] Huyck, C. (2004). "Overlapping Cell Assemblies from Correlators". *Neurocomputing* , 453-459.
- [14] Huyck, C. (2008). "CABot1: a Videogame Agent Implemented in fLIF Neurons" In *proceedings of 2008 7th IEEE international Conference on Cybernetic Intelligent Systems*, (pp. 115-120). London.
- [15] Huyck, C., & Byrne, E. (2009). "CABot1: Technical Report". Middlesex University.
- [16] Jamshed, F. and Huyck, C. (2009). "Grounding Symbols: Labelling and Resolving Pronoun Resolution with fLIF Neurons" *International Conference on Machine Learning and Applications*.
- [17] Langacker, R. (1987). *Foundations of Cognitive Grammar Vol. 1*. Stanford, CA: Stanford University Press.
- [18] Steels, L. (2007). "The Symbol Grounding Problem has been solved. So what's next." In M. deVega, *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford University Press.
- [19] Roy, D. (2003). "Grounded Spoken Language Acquisition: Experiments in Word Learning". *IEEE Transactions on Multimedia*, 5(2): 197-209.
- [20] Roy, D. and Reiter, E. (2005). "Connecting language to the world" *Artif. Intell. Vol. 167*. 1—12.
- [21] Yu, C. and Ballard, D. (2004). "A Multimodal Learning Interface for Grounding Spoken Language in Sensory Perceptions", *ACM Transactions on Applied Perception*, 1, 57-80.
- [22] Yu, C. and Ballard, D. H. (2004). "On the integration of grounding language and learning objects" In *Proceedings of the 19th National Conference on Artificial intelligence*. AAAI Press / The MIT Press, 488-493.
- [23] Belavkin, R. and Huyck, C. (2009). "A Model of Probability Matching in a Two-Choice Task Based on Stochastic Control of Learning in Neural Cell-Assemblies" In *Proceedings of the Ninth International Conference on Cognitive Modelling*.
- [24] Ostler, N, Atkins.B. (1991). "Predictable Meaning Shift: Some Linguistic Properties of Lexical Implication Rules" *Proceedings of*

the First SIGLEX Workshop on Lexical Semantics and Knowledge Representation, Springer-Verlag.

# The Design-Based Approach to the Study of Mind (in humans, other animals, and machines)

## Including the Study of Behaviour Involving Mental Processes

Aaron Sloman<sup>1</sup>

<http://www.cs.bham.ac.uk/~axs>

**Abstract.** There is much work in AI that is inspired by natural intelligence, whether in humans, other animals or evolutionary processes. In most of that work the main aim is to solve some practical problem, whether the design of useful robots, planning/scheduling systems, natural language interfaces, medical diagnosis systems or others. Since the beginning of AI there has also been an interest in the scientific study of intelligence, including general principles relevant to the design of machines with various sorts of intelligence, whether biologically inspired or not. The first explicit champion of that approach to AI was John McCarthy, though many others have contributed, explicitly or implicitly, including Alan Turing, Herbert Simon, Marvin Minsky, Ada Lovelace a century earlier, and others. A third kind of interest in AI, which is at least as, old, and arguably older, is concerned with attempting to search for explanations of how biological systems work, including humans, where the explanations are sufficiently deep and detailed to be capable of inspiring working designs. That design-based attempt to understand natural intelligence, in part by analysing requirements for replicating it, is partly like and partly unlike the older mathematics-based attempt to understand physical phenomena, insofar as there is no requirement for an adequate mathematical model to be capable of *replicating* the phenomena to be explained: Newton's equations did not produce a new solar system, though they helped to explain and predict observed behaviours in the old one. This paper attempts to explain some of the main features of the design-based approach to understanding natural intelligence, many of them already well known, though not all. The design based approach makes heavy use of what we have learnt about computation since Ada Lovelace. But it should not be restricted to forms of computation that we already understand and which can be implemented on modern computers. We need an open mind as to what sorts of information-processing systems can exist and which varieties were produced by biological evolution.

**Keywords:**

Design, niche, virtual machinery, physical, mental, model, explanation, biology, AI, commonsense.

### 1 How to describe humans and other animals

When scientists attempt to explain observations of behaviour in humans and other animals, they often use language that evolved for informal discourse among people engaged in every day social interaction, like this:

- What does the infant/child/adult/chimp/crow (etc) perceive/understand/learn/intend (etc)?
- What is he/she/it conscious of?
- What does he/she/it experience/enjoy/desire?
- What does he/she/it find interesting/boring?
- What is he/she/it attending to?
- What kinds of events surprise him/her/it ?
- Why did he/she/it do X, start Xing, stop Xing, speed up Xing... ?
- Does he/she/it know that ...?
- What did/does he/she/it expect will happen, if...?

Similar comments can be made about the terminology used in many philosophical discussions about minds, cognition, language, and the relationships between evolution and learning.

These forms of description and explanation treat the whole person or animal (as opposed to functional sub-systems) as the subject of all the verbs (of perceiving, doing, thinking, feeling, deciding, etc.). They make use of a collection of theoretical assumptions and strategies similar to what Dennett called "the intentional stance" and Newell called "the knowledge level" ([4, 14] the differences need not concern us now). That "whole animal" approach treats all those whose behaviour is being explained, whether animals, infants, toddlers, and in some cases people with serious psychiatric disorders, as if they were all basically like normal human adults in the way they operate, taking decisions and acting on the basis of what they know, what they perceive, what concepts they have, what goals, preferences and attitudes they have, and how they reason, deliberate and plan. Sometimes we can also invoke ways of being irrational, for example when experiencing strong and disastrous emotions, though it is not clear that that would be included in Dennett's "Intentional stance".

There is nothing wrong with such modes of expression if the aim is to entertain, speculate, educate in a general way, generate interest, make excuses, gossip, influence the behaviour of others, or write novels or plays. However, a different approach is needed if the aim is to provide *scientific* understanding: the kind of understanding of how humans and other animals work that could enable us to explain what they can and cannot do, how they learn and develop, or how their development can go wrong, and if we wish to gain insights into how they evolved, the relationships between evolution and development, and how deliberate external intervention can influence the processes (e.g. educational or therapeutic strategies).

Such scientific understanding is also necessary if we wish to adopt good, reliable, strategies for educating children and

---

<sup>1</sup> University of Birmingham, UK, email: a.sloman@cs.bham.ac.uk

helping people badly affected by mental abnormalities. Otherwise politicians, educators, therapists and the general public risk being influenced too much by transient fashions and misleading evidence, e.g. evidence based on statistical correlations rather than understanding of mechanisms. (There are “romantic” objections to this approach, based on a dislike of mechanistic explanations, computers, reductionism, or removal of mystery concerning human minds, e.g. [25]. This is not the place to deal with such objections, though they need to be countered.)

## 2 What is the design-based approach (“designer-stance”)?

Going beyond “common sense” descriptions and “correlational” explanations of animal<sup>2</sup> behaviours and competences requires us to formulate theories about the *mechanisms* within the animal that play a role in producing the behaviours, or that produce the competences and dispositions that produce the behaviours, just as explaining how a clock works requires us to identify components which do things that contribute to the clock’s functionality, e.g. providing the energy to keep it going, controlling its speed, detecting when to chime, or turn on an alarm sound, etc.

If we are explaining the behaviour of car or clock, we think the parts that are relevant to explaining what happens are physical components that can be identified separately from other components, and which do specific things they were designed to do. If we wish to explain what happens when a volcano erupts, or chemicals react, or a plant grows we also refer to interacting physical parts, though without assuming they were designed by humans or any other intelligent designer to do to anything.

That strategy of explaining in terms of interacting physical parts is very successful in the physical sciences, and in many branches of engineering (including explaining malfunctions in machinery as well as how things work). Physicists and chemists have learnt a lot about the items involved in physical and chemical interactions, and engineers often know a lot about what the parts of the machines they build can do in various circumstances. Many researchers hope that if only we study physical mechanisms and their connections in brains we shall achieve theories with similar explanatory power. So there is a strong temptation to look for physical parts to explain human and animal competences and behaviours, and typically that involves trying to find which bits of brains are relevant, along with which bits of bodies (sensors and effectors).

However brains are far more complex and obscure in their operation than most of the complex systems studied by physicists and chemists or the machines designed (so far) by engineers. It isn’t even clear what most of the functioning components of brains are or what their functions are, though large numbers of fragmentary discoveries are being made about which bits seem to be involved in which processes, and about how the parts interact physically and chemically.<sup>3</sup>

So, on the one hand we get neuroscientists describing components whose ability to produce processes like perceiving, deciding, learning, hypothesising, planning, wanting, evaluating is doubtful and unproven, and on the other hand we get behavioural and cognitive scientists, and even AI theorists, listing hypothesised components that are often described using familiar common sense concepts, like *perceiving, deciding, learning, ... evaluating* where

(a) nobody knows how brain mechanisms could constitute such components and (b) the concepts are too loosely defined for use in a scientific explanatory context (though not for ordinary conversation). The use of such concepts in explanations is often *circular* because the concepts presuppose that these systems have capabilities of the sorts we want to explain.

It’s as if someone tried to explain how a car engine works by listing and labelling parts, without indicating how any of the parts work or how they interact, e.g. by saying, this is the bit that starts the car, this is the bit that makes the car go faster, this is the bit that makes the car slow down, etc., leaving to others the task of specifying more precisely what exactly the parts do and explaining how they do it.

Unfortunately, when AI researchers meet the challenge by trying to specify in constructive terms how the components of intelligent systems (sensory interpreters, memory makers of various kinds, planners, choosers, emotion components, etc.) could be built, they sometimes end up using familiar labels (e.g. “emotion”, “perception”) for components whose functionality is at best a tiny fragment of what the pre-scientific uses of the labels imply – e.g. robots described as having emotions because they can smile, shake or nod their heads, etc., or described as learning because they change associative weights, or modify rules or databases. McDermott strongly criticised similar tendencies in early AI theorists [8]. The tendency re-emerges with each new wave of fashion in AI.

There have recently been attempts at trying to get beyond this vagueness and circularity by giving the robots physical bodies and sensors or motors that match some features of the biological examples. But that emphasis on embodiment is often too closely tied to attempts to replicate the rather gross morphology of the organisms, both ignoring details like the number of sensors in a mammal’s limbs, tongue, lips, etc., and ignoring what humans are able to do if their limbs and some sensors are damaged or missing, like humans born blind, or lacking limbs [22].

## 3 What sorts of mechanism do we need?

For traditional clocks, all the parts are physically separable parts, whereas we have learnt in the last half century that in the case of complex information-processing systems we need to be able to refer to more abstract concurrently active parts, namely, information-processing sub-systems i.e. pieces of so-called “virtual machinery” that are not physical, though their existence and their operation depends crucially on physical components. There need not be a one-to-one mapping between the VM components and the physical parts. (For more on the complex web of relationships between running virtual machines and the underlying physical machines see [19].)

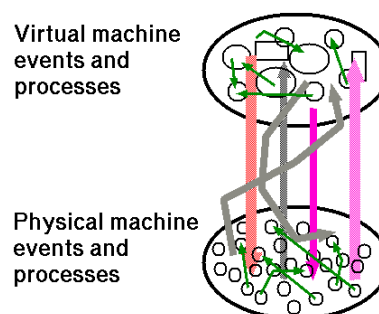


**Figure 1.** Difference in grain and causal powers between virtual machinery and physical machinery. We also need multiple layers of relative virtuality.

<sup>2</sup> By default read “animal” as including “humans.”

<sup>3</sup> See <http://news.bbc.co.uk/1/hi/7443534.stm> reporting recent research on the previously unsuspected complexity of individual synapses in mammals.

These pieces of virtual machinery are sometimes confused with programs: but programs whether printed on paper or recorded in a computer's memory, are static structures that do nothing, on their own. Something quite different comes into existence when programs run: active, enduring processes in which things get done. For example, games are played, problems are solved, machines are controlled, sensory input is interpreted, things are learnt, goals are considered and selected, plans are constructed, other processes are initiated, suspended, terminated, or in some way modified, e.g. on the basis of *information* about their progress or new developments in the environment.

The connection between processes and programs is not simple, since different processes can use the same program, invoked with different parameters, or running in different contexts – e.g. simultaneously sorting lists of different kinds.

When all that happens, there are very many physical changes in the computer scattered around the memory, the long term file store, registers in the central computer(s), switches in attached interfaces to various external devices, and so on. But the physical processes are not the only processes: things go on in computers (but not *physically inside them* as the transistors are) that are best described in a different language from the language(s) of the physical sciences: e.g. a running virtual machine can consider certain options, investigate some of them, discover that some actions lead to threats and dangers, modify or reorder its goals and preferences, and perhaps find one action that either wins the game or moves closer towards winning. Molecules, rocks, planets, electrical circuits, and, arguably, neurons cannot do those things although processes running on electrical circuits can and presumably processes running on neural, and perhaps sub-neural, mechanisms can also.

AI research has shown how to produce running virtual machines that merit such “high level” descriptions (at least to a first approximation) because of the more detailed descriptions we can give of how they acquire, manipulate, analyse, combine, derive and use various kinds of information and the (internal and external) dispositions they produce and maintain, for example, a disposition to modify a plan to achieve one or more goals in a different way, or to achieve a different goal, if new information turns up during execution.

Some philosophers argue about what is or is not “constitutive” of having an experience or a plan. When we understand, as a result of AI explorations, the space of possible designs and the variety of states and processes that can occur in each design, the importance of much discussed philosophical problems of finding necessary or “constitutive” conditions for something being an example of some pre-theoretic concept becomes utterly insignificant in comparison with the task of exploring that variety, and understanding the implications of the differences. For example, analysing, or experimenting with, working designs to find out which ones can produce systems that have various kinds of experience replaces disputes about *binary* divisions with far more useful analysis of a rich variety of cases.

Understanding diversity in a uniform way is a core goal of biology, and until now Darwinian evolutionary theory (and its more recent ramifications) provides the major example. However a different, but related study of diversity is required for understanding the variety of control systems and the variety of states and processes that can occur in them, including all the things we currently loosely label perception, learning, wanting, experiencing, deciding, disliking, etc.

## 4 Conceptual tools for explanation designers

Producing such working models depends on, among other things, constructing *information-bearers* of various kinds that express information (for the system using them), and mechanisms for operating on them, for combining, analysing, comparing, storing, searching for and using the information. These information-bearers at the lowest level of virtual machinery in computers are bit-patterns, but at higher levels can be lists of symbols, trees, graphs, arrays, equations, linked collections of changing numerals, sets of rules, running processes, and many more. Moreover these information-bearers, henceforth called “representations”,<sup>4</sup> can have different roles in different contexts: the same representation (e.g. a logical expression) could be taken to describe a state of affairs that exists, a state of affairs to be brought about, a state of affairs to be prevented, or a goal to determine whether the state of affairs exists or not. In other words, the operation of virtual machinery in computers is concerned both with semantic content and with control functions of many kinds of representation.

The notion that representations “stand for” or “stand in for” what they represent (e.g. see [24]) is a serious error, since the uses of representations are typically very different from the uses of the things they represent. You cannot climb, or paint, information about the Eiffel Tower. For more on what information is and what representations are, see [20].

At first sight, the chemical and physical structures and processes in brains do not seem to have the power to produce mental states and processes, just as transistors, metallic connections, spinning magnetic discs and the like in themselves do not seem to have the power to play chess, correct spelling, produce goals, beliefs, or plans, nor to take decisions, control inferences, make recommendations, etc. In the latter case (computing systems) we have recently learnt how to give them the powers they seem intrinsically incapable of having (though it is not a simple matter at all to do this: as half a century of research and development has shown – in particular we need far more complex systems than Turing machines). Presumably we shall eventually also learn how the far more complex physical components found in brains (see Note 3) can also be used to support a disparate range of representations and mechanisms, with many different semantic and control functions.

**NB:** At present, we don't know how neural mechanisms can encode information about complex changing and enduring 3-D structures and relations in the environment, or differences between different kinds of “stuff” that a child learns about, or causal interactions between structures in the environment, or the information used in logical and mathematical processes of reasoning. We also don't know how neural mechanisms represent “meta-semantic” information about things having thoughts, percepts, preferences, emotions, intentions, or information containing gaps. (For more information on the problems, if not the solutions see [23].)

Making progress will require us to develop a deep understanding of the intermediate-level virtual machine functionality that specifies what sorts of information processing is going on in brains, before we can produce good theories as to how the physical components do it. Just looking at the physical and chemical structures and processes found in brains, or even tracing their connections and patterns of activity, can fail to inform us as to what the higher level functions are.

So one important way in which AI may inform biological research

<sup>4</sup> This seems to be the most common use of the word “representation” in scientific contexts.

is by suggesting kinds of intermediate level information-processing functionality that may, one day, suffice to explain the observed competences (and incompetences) in various sorts of animals. Using ideas about such functionality to produce workable explanations at that level may then drive further research into how such mechanisms could be implemented in neural systems, and also research into ways in which they can go wrong, how they develop in individuals, how they differ across species, and how they might have evolved.

Making advances that increase explanatory power of our theories at the intermediate levels is very hard, however – in part, because it is hard to identify the requirements to be met, as decades of oversimplified goals adopted by AI researchers have shown.

## 5 Whole organism explanations

Often, scientists (or philosophers) attempt to produce explanations of the phenomena observed, or hypothesised, by describing what is going on *inside* the person or animal, but the ways they have of doing that derive from concepts used in everyday conversation for describing human mental states and processes, such as *noticing, seeing, expecting, deciding, comparing, choosing, learning, hypothesising, wanting, preferring*, and many more. One characteristic of the above concepts is that they normally refer to what a whole person is doing, e.g. it is you that notices something, not some portion of your brain, or mind, or one of your eyes or ears. What we need are more fine-grained process descriptions, but with the power to explain, and also contradict and refine, our more coarse-grained common sense descriptions. (The corrections would be analogous to corrections of everyday categories in that occurred in previous branches of science: e.g. there isn't a kind of "bad air" that causes illness and whales are not fish.)

Many scientists have tried to avoid those common-sense categories of states, actions and processes, when constructing explanatory theories. Some attempted to use only concepts defined in terms of observable behaviour with consequences criticised by (among others) Chomsky in [2]. Others, as explained above, try instead to refer to *physical* parts of a person or animal which are assumed to have the required explanatory power. (So-called "mirror neurons" illustrate this move).

A third group refer to *hypothesised* non-physical parts labelled in terms of their cognitive functions, and sometimes represented in diagrams of boxes and arrows, whose operations are supposed to explain what happens. Often there is no specification of how they produce those operations, how to make the components, how to test what they do, or how to vary them, unlike parts of a grandfather clock, which can be made in a factory, given changeable weights or linked in various ways with other parts, and unlike parts of an information-processing system assembled from various hardware and software computing mechanisms that can be combined in different ways to produce different competences.

## 6 A designer's explanatory requirements

We can build a clock that conforms to a suggested explanation and see if it works like the one that puzzled us – empirical testing. Or we can use our knowledge of geometry, mechanics, dynamics and mathematics to work out what behaviours the proposed structure could produce – analytical/mathematical testing of a theory. But mostly we cannot do either for commonly hypothesised parts of minds, because the specifications are generally too loose – e.g. *being surprised, attending, noticing, wanting, learning, remembering*, etc.

If all we can say to describe hypothesised parts is that they explain the phenomena they were constructed to explain (this bit does seeing, that bit does deciding, another bit does learning, ....) then it is not clear that we are explaining anything.

However, there is another alternative to explanations in terms of categories from behaviourism, physics, chemistry or physiology, that avoids such explanatory vagueness, indicated in previous sections. As a result of decades of research in AI, building on research of other kinds that have steadily extended the powers of computing systems, we have begun to learn how to give explanations of mental processes, and the behaviours they produce, in terms of *information-processing* mechanisms and architectures that are described at a higher level of abstraction than brain mechanisms and at a lower level than common-sense descriptions.

So the vagueness characteristic of many psychological explanations can be reduced, and explanatory power can be increased, by using parts at a lower level of information-processing machinery that we have previously demonstrated in working systems can do the sort of thing we are talking about. This is like going from kinds of matter (water, air, salt, mud, wood, etc.) to kinds of molecules in the history of physics and chemistry. An important feature of this approach is that it assumes that the intermediate level mental mechanisms, states, events and processes are not just useful fictions (as some of Dennett's wording suggests), but actually do *cause* things to happen, including other mental processes and also physical processes, such as changes in brain states and external actions, just as software engineers assume, with good reason, that the calculations and rule applications that go in a computer system can cause both other calculations and rule applications, and also changes in internal physical states as well as on external computer screens and other attached physical devices. This contrasts, for example, with Dennett's "intentional stance" [4] and with "mind-brain identity" theories<sup>5</sup>. For more on the causal states of virtual machinery see [19].

## 7 Developing a better explanatory ontology

What sorts of parts/components should be referred to in adequate explanations is not easy to understand – and unfortunately there are ill-judged fashions at work. As explained in Section 3, there is a mode of explanation of how complex systems work that is very different *both* from describing their physical, chemical, electrical, or mechanical parts and operations *and* from describing them using common mentalistic language. It is based on what we have learnt about designing complex information processing systems of many kinds, none of which come near the specific kinds of sophistication that we wish to explain in humans (young and old) and other more or less intelligent animals, though there are promising signs of progress, as computer-based systems do more and more things that previously could only be done by humans and other animals.<sup>6</sup> The particular forms of explanation that have been developed refer to such things as

- The kinds of information that the organism acquires and uses.
- How the information is acquired (including which features of the environment make it available and how the sensory and perceptual mechanisms acquire it).

<sup>5</sup> <http://plato.stanford.edu/entries/mind-identity/>

<sup>6</sup> A particularly impressive robot with a subset of insect-like intelligence is Boston Dynamic's BigDog <http://en.wikipedia.org/wiki/BigDog>.

- The various ways in which the information can be manipulated, analysed, recombined, derived, or used in planning and problem solving.
- The forms of representation used to encode that information and the mechanisms that operate on those forms of information.
- The ontologies that constitute the basic information structures used, from which more complex information is constructable.
- The architectures in which those various capabilities, mechanisms, and information structures are combined.
- How different processes can run concurrently within an architecture, with or without conflicts, and how conflicts can arise, how they can be detected and how they can (in some cases) resolved.
- What sorts of self-monitoring and self-control mechanisms can operate, monitoring and controlling different subsystems.
- How all the items listed above (kinds of information, forms of representation, mechanisms, architectures, etc.) are initially constructed and how they continue to grow and develop over extended periods in some organisms (e.g. humans).
- The kinds of information, forms of representation, mechanisms and architectures that need to exist at a very early stage to support (bootstrap) all those developments.
- The ways in which the nature of the environment constrains the types of information that are available to the organisms and poses problems that the information needs to be used to solve. (For two different species living in the same location, the role of the environment can be very different, because of the effects of their different evolutionary past, producing different bodily forms, different needs, different goals, different forms of information-processing, different forms of reproduction, etc.)

In talking about a mind we are talking about a complex system with many concurrently active parts, whose workings need to be explained in terms that can help to bridge the gap between their functions and the underlying physical mechanisms that make it possible to have such functions and which limit and shape those functions. We also need an explanation of how those parts interact, harmoniously most of the time, and what happens when they come into conflict.

These parts are organised in an information-processing architecture that maps onto brain mechanisms in complex, indirect ways that are not well understood.

So, when studying some human (or animal) psychological capability or limitation, we should ask questions like this if we wish to acquire a deep scientific understanding (as opposed to close human-to-human empathetic understanding for example):

- Which parts of the information-processing architecture are involved in the capability or deficiency?
- What are their functions?
- What kinds of information do they acquire and use?
- How do they do this?
- What is the total architecture in which the various parts function?
- How is the information represented?  
(It could be represented differently in different sub-systems for different purposes).
- What kinds of manipulations and uses of the information occur?
- What mechanisms make those processes possible?
- How are the internal and external behaviours selected/controlled/modulated/coordinated?

- How many different virtual machine levels are involved and how are they related (e.g. physical, chemical, neural, subsymbolic, symbolic, cognitive,...)?

For example, a parrot can use one foot to balance on a perch while at the same time alternately holding a walnut in its beak or the other foot, as it rotates it trying to find a good place to bite into it. A particular robot could be built that does nothing but produce that behaviour, though only with a very specific shape of perch and a very limited range of sizes and shapes of walnut. But a robot built using contemporary AI techniques would not be able to *work out* how to do such things starting from a repertoire of competences that could be used for a wide variety of purposes, e.g. alternately using beak and feet while climbing. (Some of the problems, but not all, are due to limitations of current mechanical and electronic engineering, including power weight ratios, strength weight ratios, sensor limitations, etc.)

## 8 Limitations and common errors in current AI

Of course, where current AI researchers and their models use flawed assumptions, theories or designs, biologists need to resist being inspired by them! For example, the switch from common-sense descriptions to descriptions that have been useful in designing working computational models is not always an improvement, since sometimes the ontology available to AI designers is too restrictive.

A simple example concerns the type of programming language used. Some programming languages, especially the earliest ones, were aimed almost exclusively at specifying numerical computations, whereas from the beginnings of AI it was clear that computers would have to manipulate non-numerical information structures, including sentences, logical expressions, grammars, parse trees, plans, equations, and various structures built from those. AI languages like Lisp, Pop-11, Prolog, Scheme and others were designed accordingly.

However there are many AI/Robotics researchers who are unfamiliar with such languages and whose programming skills are mostly geared to numerical computations. As a result such designers develop systems in which all information about sensory input, about structures and processes in the environment, about motivation, and about control of actions is expressed in numerical form, including, for example, the use of a global cartesian coordinate system to represent spatial locations, orientations, distances, sizes, and relationships – a practice criticised in this presentation [18].

An issue that some AI designers do not face up to is whether to think of all software components of an intelligent system as running on a single processor, or as using multiple concurrently active processors. Very often the only development environment available to AI researchers assumes either one or a very small number of processors, and uses programming technology that does not easily support development of asynchronous concurrently active subsystems. A result of this is that such researchers have to work on techniques for interleaving different activities, and strategies for optimising the allocation of processing resources between them. An example is a great deal of work involving addition of metacognitive capabilities to intelligent systems, which has addressed the need to optimise the sharing of resources between cognitive and metacognitive processes, and possibly other processes.

All that effort is wasted in the context of modelling natural cognition, if brains have different subsystems running in parallel performing the different tasks. An extreme version of the error

is the widely accepted assumption that intelligent robots need to make use of a repeated cycle of cognitive phases such as this:

**sense→think→decide→act**

In contrast, our own work has assumed that many processing components of biologically inspired robot can, and need to, run in parallel, as indicated in our work on the CogAff architecture scheme and its H-Cogaff special case (<http://www.cs.bham.ac.uk/research/projects/cogaff/#schema>).

Also bad theories in AI, e.g.

Symbol grounding theory

theories about emotions

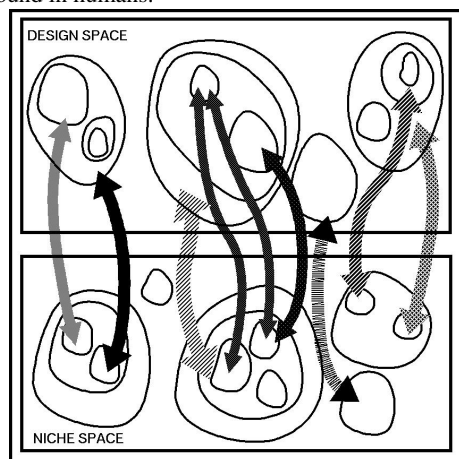
theories that all motivation has to be reward based.

Another restrictive influence can be the assumption that existing information processing architectures or other existing techniques are adequate to the task of building explanatory models, when what is badly needed is new deep requirements analysis: examining the biological phenomena (e.g. animal behaviours, or studies of brain mechanisms) with very great care to see whether something can be learnt that improves the way in which AI can serve biology.

## 9 Different requirements and different designs

It is important that in studying those questions we remember that humans (and other animals) evolved to function in a particular type of environment found on this planet. Moreover there are difficult solutions to different subsets of the problems of surviving in such an environment. A deep understanding of one design would include being able to compare and contrast it with other designs and with different sets of requirements and constraints imposed by the environment (including other animals).

Much writing in philosophy and psychology falls into the trap of assuming there is only one way of perceiving, wanting, believing, deciding, and being conscious – the human way. In trying to describe the human way they often make the mistake of forgetting that we are products of evolution and there were many different distinct designs in our precursors. If we can find theoretical or empirical evidence for some *alternative possible* requirements and designs, we'll be in a far better position to understand the costs and benefits of the *actual designs* found in humans.



We need to understand different sets of requirements (niches) different designs, and the many complex ways designs can relate to requirements (not just in terms of numerical fitness functions). We also need to understand the many kinds of trajectories through niche space and design space that can occur in evolution, in individual development, in cultural evolution, and so on. Dennett's little book

on "Kinds of minds" [5] illustrates this, but using very "broad brush" categories.

## 10 Kinds of requirements

When presenting theories about cognitive functions and the mechanisms that explain them it is very important to try to be clear about the precise collection of requirements which the proposed mechanisms are supposed to meet. In particular, we need to distinguish at least the following, though far more fine-grained distinctions between requirements will be needed.

- Producing behaviour in real time that is suited to the precise configurations of things in the environment that define the goals, and the positive and negative affordances.
- Thinking, reasoning, explaining or making plans concerning actions that are not currently being performed but which could be performed in the future, or were performed in the past (by the person or animal concerned).
- Perceiving thinking, reasoning about, perceived processes occurring in the environment not caused by the individual, but which may or may not affect the individual (proto-affordances), or may be relevant to the goals or actions of another individual (vicarious affordances).
- Perceiving thinking, reasoning about, the percepts, thoughts, desires, plans and actions of another intelligent agent, as opposed to something like wind, water or gravity that can cause things to happen without using any cognitive mechanisms. Being able to perceive, think, reason, or deliberate about other individuals with similar powers requires **meta-semantic competences**, which not all organisms seem to have.
- Being able to use "self-directed" meta-semantic competences applied to one's own thinking, reasoning, perceiving, etc., for instance finding a flaw in one's planning strategy and repairing it. Can it recurse?

## 11 Kinds of Observations Needed

A task on which more thought is required is how the research goals listed here can influence choice of biological experiments and the observations required. (See the paper by Chappell and Thorpe in these proceedings.)

One implication that is rarely noticed is that insofar as different individuals have different combinations of knowledge, concepts, forms of representation and possibly also architectures (e.g. if they are at different stages of development), important information may be lost by focusing on averages across collections of experimental subjects, as opposed to adopting a "clinical" approach and trying to describe in detail what exactly different individuals do and what that implies regarding differences in how they do things.

This can also be important in making studies of development more fine-grained than is common when averages across populations in a species or in an age group are used, often without even paying attention to the variance! (I see many presentations where graphs show changes in averages without showing any of the variance, alas, indicating a deep flaw in our education of scientists.)

## 12 A consequence of adopting this approach

If we think of features of humans and other animals such as consciousness, intelligence, attention, memory, emotions, autonomy

in this ‘design-based’ way (adopting what John McCarthy now calls ‘the designer stance’, in [7]) the sorts of questions we can ask and the sorts of theories we can consider are expanded in an important way.

A design for a working system (microbe, ant, chimpanzee, human, robot) will specify a complex virtual machine with many coexisting, interacting information-processing components.<sup>7</sup>

Since there are many components, it is possible to consider different designs for working system that use different combinations of such components, and different versions of the components. This sort of variation in designs is evident in the products of biological evolution.

A corollary is that where we are naturally inclined to think of a *binary* division such as a division between animals that do and do not have some feature X (consciousness, creativity, autonomy, emotions, planning capability, free-will, etc.) the design based approach replaces the binary division by much richer spaces of possibilities, including *taxonomies* or *generative grammars*. We should not assume that the only alternative to a binary division is a linear continuum (differences of degree). Biological changes, insofar as they are based on molecular changes and other structural changes (e.g. duplication of a component) are inherently *discontinuous*. Our theories of the spaces involved must accommodate this.

### 13 An example, from AI

Many people (including the author years ago) assume that there is a binary division between *reactive* and *deliberative* control mechanisms. After hearing several presentations and reading several documents making use of these labels in confusingly different ways, I eventually realised that people were interpreting the division in different ways because the space of possible designs had a kind of complexity that had not been studied properly and people were basing the distinction on different ‘cracks’ in the space.

For example, some people were using ‘deliberative’ to refer to any system that could, in some sense, evaluate alternative action possibilities and select one, whereas others used the label to refer to more complex systems that can plan more than one step ahead when taking decisions.

When I looked closely, I found that there were several more important sub-divisions between different sorts of deliberative competence, and documented them in [21]. I don’t claim that the analysis of that document is complete: there are important sub-cases to be distinguished, especially if we are to understand the stages in the evolution of the more complex competences.

When considering any competence or mechanism of type X and asking which animals have X, how X evolved, what X’s costs are, what X’s benefits are, which neural or other mechanisms are involved in X, etc. a good heuristic is to ask

- How many varieties of X are there?
- what sorts of distinct components, that might have evolved separately, are involved in different varieties of X?

And that generally has the result of replacing a binary divide between things that do and do not have X with a sometimes large collection of varieties of X, and a large collection of intermediate cases between not having a particular sort of X and having it.

This idea was used in an analysis of the notion of ‘free will’ originally posted to a ‘usenet’ news group, now available in [15]. Similar ideas are in [3] and [6].

### 14 Relevance to ontology/ontologies

The previous section pointed out a consequence of this study of varieties of architecture that might have been designed or might have evolved in response to explicit or implicit requirements such as the pressures of an ecological niche, namely that we usually need a richer ontology of types of design than can be expressed using binary distinctions normally assumed.

Another consequence of the approach is that the process of designing a complex working architecture, testing it and finding problems that need to be addressed by improvements in the design, often teaches us that there is a much richer variety of possible internal states than we might have considered possible in advance.

Systems with complex virtual machines [19] that include concurrently active interacting sub-systems, including some sub-systems that monitor and control others, can have a richer variety of internal states and processes than can be defined in terms of varieties of external behaviour, or even relations between inputs and outputs. For example, a system can run internally and have no connections with output signals most of the time, even though it occasionally is linked to inputs and outputs.

A simple example could be a complex virtual machine that is capable of playing many different games, and at any time practices some of those games internally by playing itself, e.g. at chess, or draughts (checkers), as a result of which its competence in those games increases, though there is no external sign of those changes unless it engages in a game with an external player, which may never actually happen.

Its input and output channels may have capacity limits that limit the total number of games that the system actually plays in its lifetime, and that limit may be significantly lower than the number of different games it has the competence to play.

By studying the variety of internal states that the architectural design (the information-processing architecture) of some organism makes possible we may find that to understand and explain how the organism works we need a much richer ontology of states and processes than would be suggested merely by watching its behaviours and trying to classify them.

This is particularly rue of humans: there is no reason to suppose that the ontology expressed in our ordinary language concepts for talking about mental processes, or even the extensions to that ontology developed by psychologists and psychiatrists as a result of interacting with and experimenting on humans is rich enough to account for all the important phenomena of human life: instead we need a much richer ontology of states and processes derived from a good theory of how the system works. This is similar to the way our understanding of the variety of types of material substance had to be substantially revised when we discovered the underlying architecture of matter, as composed of atoms of various sorts that can combine to form molecules of various sorts that can be arranged in configurations of various sorts – none of which was dreamt of prior to the development of modern physics and chemistry.

<sup>7</sup> As explained in this talk on “Why robot designers need to be philosophers – and vice versa”  
<http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk51>

## 15 Logical geography vs logical topography

The issues raised here are pursued further in different ways in different online papers produced as part of the Cognition and Affect project and the CoSy Robot project. One of the papers, [16], discusses relationships between philosophy and science in the context of an attempt to clarify Ryle's notion of 'Logical Geography', showing that there is a deeper type of investigation, which I called the study of 'Logical Topography', which identifies aspects of some portion of reality that allow various possible kinds of concepts to be developed, in contrast with the study of the concepts that are *actually* in use constituting Ryle's 'Logical Geography'.

The difference emerges in two ways: The study of logical geography assumes (a) that there is one collection of concepts whose relationships can be charted and (b) that this will answer philosophical questions definitively. The study of logical topography reveals (a) that the relevant aspect of reality can be divided up in different ways, leading to different logical geographies, and (b) that that reality may itself have unnoticed complexity of structure, which, when explored in depth, shows possibilities that were not exposed by the original philosophical investigations.

On the basis of those ideas, we can see that philosophical theory building has much in common with scientific theory-building (including the ability to introduce extensions to our ontology), and which uses abduction. However, for philosophers and researchers studying natural intelligence, to ignore the advances in the logical topography of information processing systems would be analogous to chemists continuing, like the old alchemists, to find and use laws of how different kinds of stuff interact without attempting to move to a deeper level of explanation, as emerged in the atomic and sub-atomic theory of matter.

## 16 Conclusion?

There is no conclusion, because we are still in the early stages of a very rich and deep exploration whose implications are potentially profoundly important not only for the biological sciences, but also for our understanding of ourselves. A revised, extended version of this paper will be written after the symposium, possibly with co-authored sections.

## ACKNOWLEDGEMENTS

I believe John McCarthy's use of the term "the designer stance" is closely related to what I have called "the design-based approach". See his paper [7]. I tried to take his ideas a bit further in [17] and the presentations here: <http://www.cs.bham.ac.uk/research/projects/cogaff/talks/> Much of Minsky's work, is in a similar vein. I think he has been moving upwards in the layers of virtual machinery in his recent work, and it is not always easy for readers to understand how the ideas might be implemented. See [9, 10, 11, 12, 13] The most exhaustive and historically deep survey of these ideas can be found in Margaret Boden's "magnum opus", [1]. Two extracts from that have been made available for this symposium.

## REFERENCES

- [1] M. A. Boden, *Mind As Machine: A history of Cognitive Science (Vols 1-2)*, Oxford University Press, Oxford, 2006.
- [2] N. Chomsky, 'Review of skinner's *Verbal Behaviour*', *Language*, **35**, 26-58, (1959).

- [3] D.C. Dennett, *Elbow Room: the varieties of free will worth wanting*, Oxford: The Clarendon Press, 1984.
- [4] D.C. Dennett, *The Intentional Stance*, MIT Press, Cambridge, MA, 1987.
- [5] D.C. Dennett, *Kinds of minds: towards an understanding of consciousness*, Weidenfeld and Nicholson, London, 1996.
- [6] Stan Franklin, *Artificial Minds*, Bradford Books, MIT Press, Cambridge, MA, 1995.
- [7] J. McCarthy, 'The well-designed child', *Artificial Intelligence*, **172**(18), 2003-2014, (2008). <http://www-formal.stanford.edu/jmc/child.html>.
- [8] D. McDermott, 'Artificial intelligence meets natural stupidity', in *Mind Design*, ed., J. Haugeland, MIT Press, Cambridge, MA, (1981).
- [9] M. L. Minsky, 'Steps towards artificial intelligence', in *Computers and Thought*, eds., E.A. Feigenbaum and J. Feldman, 406-450, McGraw-Hill, New York, (1963).
- [10] M. L. Minsky, 'Matter Mind and Models', in *Semantic Information Processing*, ed., M. L. Minsky, MIT Press, Cambridge, MA, (1968).
- [11] M. L. Minsky, *The Society of Mind*, William Heinemann Ltd., London, 1987.
- [12] M. L. Minsky, *The Emotion Machine*, Pantheon, New York, 2006.
- [13] M. Minsky, 'Interior Grounding, Reflection, and Self-Consciousness', in *Brain, Mind and Society, Proceedings of an International Conference on Brain, Mind and Society*, Tohoku University, Japan, (September 2005). Graduate School of Information Sciences, Brain, Mind and Society. <http://web.media.mit.edu/~minsky/papers/InternalGrounding.html>.
- [14] A. Newell, 'The knowledge level', *Artificial Intelligence*, **18**(1), 87-127, (1982).
- [15] A. Sloman, 'How to Dispose of the Free-Will Issue', *AISB Quarterly*, **82**, 31-32, (1992). <http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#8>.
- [16] A. Sloman, 'Two Notions Contrasted: 'Logical Geography' and 'Logical Topography' (Variations on a theme by Gilbert Ryle: The logical topography of 'Logical Geography'.', Technical Report COSY-DP-0703, School of Computer Science, University of Birmingham, Birmingham, UK, (2007). <http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0703>.
- [17] A. Sloman, 'The Well-Designed Young Mathematician', *Artificial Intelligence*, **172**(18), 2015-2034, (2008). <http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0807>.
- [18] A. Sloman, 'Ontologies for baby animals and robots. From "baby stuff" to the world of adult science: Developmental AI from a Kantian viewpoint.', 2009. Online tutorial presentation: <http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#brown>.
- [19] A. Sloman, 'Virtual Machines and the Metaphysics of Science', Sep 2009. <http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#mos09>, PDF presentation for AHRC Metaphysics of Science Conference.
- [20] A. Sloman, 'What's information, for an organism or intelligent machine? How can a machine or organism mean?', in *Information and Computation*, eds., G. Dodig-Crnkovic and M. Burgin, World Scientific, New Jersey, (To appear). <http://www.cs.bham.ac.uk/research/projects/cogaff/09.html#905>.
- [21] Aaron Sloman, 'Requirements for a Fully Deliberative Architecture (Or component of an architecture)', Research Note COSY-DP-0604, School of Computer Science, University of Birmingham, Birmingham, UK, (May 2006). <http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604>.
- [22] Aaron Sloman, 'Some Requirements for Human-like Robots: Why the recent over-emphasis on embodiment has held up progress', in *Creating Brain-like Intelligence*, eds., B. Sendhoff, E. Koerner, O. Sporns, H. Ritter, and K. Doya, 248-277, Springer-Verlag, Berlin, (2009). <http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0804>.
- [23] A. Trehub, *The Cognitive Brain*, MIT Press, Cambridge, MA, 1991. <http://www.people.umass.edu/trehub/>.
- [24] B. Webb, 'Transformation, encoding and representation.', *Current Biology*, **16**(6), R184-R185, (2006).
- [25] J. Weizenbaum, *Computer Power and Human Reason: From Judgement to Calculation*, W.H. Freeman, San Francisco, 1976.

# What's Life Got To Do With It?<sup>1</sup>

Margaret A. Boden<sup>2</sup>

## Abstract.

This is a re-formatted version of section 16.x of M. A. Boden, *Mind As Machine: A history of Cognitive Science (Vols 1–2)*, Oxford University Press, Oxford, 2006. It discusses the question whether mind can occur without life.

The book has a large bibliography. All the references in this extract are to items listed in the bibliography.

It is included in the AIIB proceedings and web page with the kind permission of the author. Letters in the section headings refer to sub-sections in the book.

## 1 Introduction

All the minds we know about are found in living things. But why?

- Couldn't there be mind and meaning without life?
- And what is life?
- Given that it involves self-organization, just what sort of self-organization is it?
- Must it involve evolution, for instance?
- Is embodiment essential? And what's that? Is mere physicality enough for embodiment?
- Or is metabolism needed too—and, again, what's that?
- What's the link, if any, between metabolism and mind?
- Could life be generated artificially, and if not why not?
- Is strong (i.e. virtual) A-Life impossible in principle?
- If so, does it follow that strong AI is an illusion too?

These questions took a long time to surface in cognitive science. Or perhaps one should rather say to *resurface*. For in the very early days, life and mind were both discussed—and were treated largely on a par.

The cyberneticians of the 1940s applied their theories of self-controlling machines to both living organization and purpose (see 4.v–vii). Life wasn't regarded as necessary for teleology, for self-guided missiles were said to exemplify goal seeking (Rosenblueth, Wiener and Bigelow 1943). However, since life and mind were supposed to consist in fundamentally similar principles of control, it didn't seem surprising that all the minds we know about are grounded in living things.

McCulloch had been a member of this life-and-mind movement since its inception in the 1930s. But he'd been deeply interested in logical analyses of language even before then (see 4.iii.c). In the event, his paper of 1943 turned attention away from life in favour of mind.

Control by feedback gave way to logic-based computation. The NewFAI computer-modelling community saw mind and meaning as matters for propositional logic, their origin in adaptive behaviour and living embodiment being downplayed. Even unreconstructed cyberneticians now sometimes spoke of mind without mentioning life: when William Ross Ashby wrote a paper on cybernetics for *Mind*, he concentrated on defending materialism against mind-body dualism, not on exploring the philosophy of self-organization (Ashby 1947).

In short, by mid-century the link between life and mind—though still widely accepted, even taken for granted—wasn't explicitly considered. The people on the NewFAI side of the emerging cybernetics/symbolic schism (4.ix) ignored life and spoke only of mind. The people on the other side said very little about mental phenomena beyond perception and goal-seeking: self-reflection and reasoning were mostly ignored.

That life/mind split within cognitive science lasted for several decades. Much the same was true in philosophy, especially the analytic variety. It's still the case, in 2005, that the link between life and mind is ignored (beyond mere lip-service) by the mainstream. However, these issues are now arousing interest—largely as a result of the rise of A-Life.

## 2 a: Life in the background

Even in the early 1960s, there were a few exceptions to what I've just said. Most important, with respect to their historical role in cognitive science, were the people impressed by cybernetics who were developing holistic philosophies—and computer models—of life and/or mind (15.vi–vii).

The most influential of these (all discussed later in this Section) were to be the Chilean neuroscientist Maturana (with Varela and Milan Zeleny), whose similarity to Dreyfus was mentioned in Section vii.b above, and the philosopher Howard Pattee, with his students Robert Rosen and Peter Cariani. Both Maturana and Pattee were working on the concept of life from the 1960s. Although they had some early disciples, their ideas didn't become widely known until the 1990s. By the new millennium, however, Maturana and Varela's ideas had been published in semi-popular form, and an entire issue of the journal *BioSystems* was devoted to Pattee's work and influence (Rocha 2001).

All of the above were either scientists (like Maturana) or philosophers very close to science, so close that they sometimes got involved in scientific work (Pattee, for instance). None were “pure” philosophers.

In general, the mid-century philosophers who were interested in the puzzle of life-and-mind came from the Continental, not the analytic, side of the fence. As a result, they were largely ignored by the

<sup>1</sup> ©Margaret A. Boden 2006

<sup>2</sup> University of Sussex, UK, M.A.Boden@sussex.ac.uk

scientific community. Moreover, they were exceptions even within their own, neo-Kantian, tradition. To be sure, phenomenologists in general took the human being's embodied living-in-the-world as philosophically basic. And Wittgenstein saw language as part of our "natural history", declaring: "What has to be accepted, the given, is—so one could say—*forms of life*" (1953: 226, italics in original). Most of them had scant interest, however, in what biologists mean by life—which includes oak trees and barnacles, as well as human beings.

One exception to this was the existentialist theologian Hans Jonas (1903-1993), who developed a new philosophy of biology in the 1950s. Unlike Maturana and Varela, he wasn't interested in biology for its own sake, but as an aspect of what he saw as the disastrous cultural *denouement* of Descartes' materialism (Jonas 1966: 58-63).

An ex-pupil of Heidegger, Jonas fled Germany for England when the German Association for the Blind expelled its Jewish members (Jonas 1966: xii). From the mid 1950s to the mid 1970s he worked at the New School of Social Research, in New York. Despite still regarding Heidegger as "the most profound and ... important [proponent] of existential philosophy" (229), he rejected his philosophical dichotomy between humans and other living things—and his pro-Nazi sympathies, too. He explained the latter in terms of "the absolute formalism of [Heidegger's] philosophy of decision," in which "not *for* what or *against* what one resolves oneself, but *that* one resolves oneself becomes the signature of authentic Dasein" (Jonas 1990: 200). And that, in turn, he saw as a result of the stripping-away of value from nature, its "spiritual denudation" by Descartes and modern science (Jonas 1966: 232).

It was in response to this disenchantment of nature that Jonas published various essays on life in the post-war years, and collected them as *The Phenomenon of Life* in 1966. They outlined a framework for a biology that would admit value as an intrinsic feature of life in general. ("Outlined" and "framework" are important here: he discussed almost no specific examples.)

Embodiment, and in particular metabolism, was seen by Jonas as philosophically crucial (1966: 64-91). Not only was life essential for the emergence of mind (99-107), but *all* self-organized matter was, in a sense, ensouled—though where Maturana and Varela spoke of life as involving *cognition*, Jonas spoke of life as involving *self-concern*. (He lauded Heidegger for having "shattered the entire quasi-optical model of a primarily *cognitive* consciousness, focusing instead on the wilful, striving, feeble, and mortal ego"—1996: 44; italics in original.) As he put it:

One way of interpreting [the ascending scale of life] is in terms of scope and distinctness of experience, of rising degrees of world perception.... Another way, concurrent with the grades of perception, is in terms of progressive freedom of action.... [One] aspect of the ascending scale is that in its stages the "mirroring" of the world becomes ever more distinct and self-rewarding, beginning with the most obscure sensation somewhere on the lowest rungs of animality, even with the most elementary stimulation of organic irritability as such, in which somehow already otherness, world, and object are germinally "experienced," that is, made subjective, and responded to. [We spoke, above, of freedom.] One expects to encounter the term in the area of mind and will, and not before: but if mind is prefigured in the organic from the beginning, then freedom is. And indeed our contention is that *even metabolism, the basic level of all organic existence, exhibits it: that it is itself the first form of freedom* (Jonas 1966: 2f.; italics added).

Even in "the blind automatism of the chemistry carried on in the depths of our bodies," there is "a principle of freedom ... foreign to suns, planets, and atoms". For living organisms have a special type of identity and continuity: a stable dynamic form made of an ever-changing material substrate. In short, "mind is prefigured in organic existence as such" (5). Plants, too, have "metabolic needs", although they stand in an "immediate" relationship to their environment. And metabolism is the necessary base of all forms of mediation: perception, motility (action), emotion, and—ultimately—conscious imagination and self-reflection. (These phenomena emerge as a result of evolution: Darwin, *despite* his materialist assumptions, had enabled us to understand this: 38-58.) Life and mind are ontologically inseparable: "the organic even in its lowest forms prefigures mind, and ... mind even on its highest reaches remains part of the organic" (1).

In other words, Jonas was offering an answer to the question of *why* all the minds we know about are found in living things. At the same time, he was offering an answer to the question of *what life is*.

He explicitly refused to speculate about the origins of life (4), even though this was already being discussed by biochemists (Chapter 15.x.b). His interests were ontological, not scientific: metabolism was "the break-through of being" from mere physicality to "the indefinite range of possibilities which hence stretches to the farthest reach of subjective life ..." (3). (Accordingly, he retained a Heideggerian hostility to technological theories/analogies of life or mind: 108-126.)

The book was reissued (by several different publishers) in 1979, 1982, and 2001, and translated into German in 1994. So one can't say that it was wholly ignored. Indeed, because of his stress on the intrinsic value of life and humankind's responsibility towards it, Jonas' work—especially his volume on ethics (Jonas 1984)—has recently become better known thanks to the environmentalist movement.

In the 1950s and 1960s, however, his philosophy of biology was ignored by analytical philosophers and mainstream biologists alike. (And by cyberneticists too, whose analysis of living purpose and rocket teleology *in the very same terms* he'd rejected as "spurious and mainly verbal"—1966: 111.) The same was true of Maturana and Varela's early work, but they have now earned a clear, if still marginal, place in the history of cognitive science. Jonas has not (but see Di Paolo in press). He's relevant here not as a protagonist in that historical drama, but as a mid-century philosopher who tried to argue the case that mind requires life, rather than taking it for granted.

Another philosopher who'd done this was Henri Bergson (Chapter 2.vii.c). By the end of the twentieth century, Bergson's views on "creative evolution" were being revived in some philosophical circles—especially in "process" philosophy/theology (Sibley and Gunter 1978; Papanicolaou and Gunter 1987). This emulated Bergson alongside the even greater hero Alfred North Whitehead (4.iii.b). But some unorthodox scientists were taking an interest too. The physicist/philosopher Henri Bortoft (1996) put Bergson second only to Goethe as a precursor of current dynamical theories in science and philosophy (see 2.vii.c). And a few neo-Bergsonian philosophers even tried to relate his ideas to cognitive science and/or A-Life.

For example, Gilles Deleuze (1925–) revived certain aspects of Bergson's philosophy by stating them in terms of ideas about dynamical systems (Deleuze 1966/1988). I'm saying that at second hand, I must confess, for Deleuze himself is nigh unreadable by anyone more accustomed to analytical philosophy. Much as Richard Montague's work couldn't spread among linguists until a clear account of it had been provided by Barbara Partee (see 9.ix.c), so Deleuze's has been

made accessible to cognitive scientists by his expositor Manuel DeLanda (2002).

Although he rejected Bergson's dualist interpretation of *élan vital*, Deleuze offered a "re-enchantment" of matter, nevertheless. He even (confusingly) used the term "spirituality" in talking about matter and life. But this wasn't intended as transcendent spirituality: rather, it referred to the abstract principles of self-organization, and the structured spaces of possibilities, that are inherent in matter/energy.

He saw matter not as inert stuff subject to external influences, but as the source of formative material *processes*. A soap-bubble, for instance, actively minimizes the surface tension at every point (it dynamically "computes" its own shape). The dynamical structures generated by matter were said to be constrained, in part, by abstract topological principles describing connectivities and attractors of various kinds (compare Stuart Kauffman's work on NK networks, and Randall Beer's on CTRNs: 15.ix.b and xi.b).

On this view, life was a special case of matter, and mind a special case of life. It followed that there's no *special* difficulty about giving a naturalistic, even a materialistic, account of mind or intentionality, even though spelling one out in detail may be highly challenging.

However, these intriguing analogies weren't helpful in furthering scientific understanding. (Or anyway, they haven't been helpful yet: DeLanda's relatively accessible version of Deleuze appeared only two years ago, and it remains to be seen whether many scientists will take it up.) Admittedly, the ever-maverick neuroscientist Karl Pribram (1987)—accused in the early 1960s of actually *believing* the MGP manifesto (6.iv.c)—described the cerebral basis of some cognitive processes in Bergsonian terms. But that's not to say that he *used* Bergson's ideas to make discoveries which otherwise would not have been made. Rather, he pointed out an analogy between Bergson's views on memory and his own (longstanding) holographic/holonomic theory (cf. 12.v.c).

Cognitive scientists who weren't already sympathetic to dynamical systems and/or Kauffman's approach to A-Life weren't likely to be interested in Bergson's work at all, even if they encountered it. And that was unlikely: as remarked in Chapter 2.vii.c, it had been more or less forgotten since mid-century—especially by philosophers of an analytic cast of mind.

For over thirty years, then, the concept of life was usually ignored in discussions of mind as machine. To be sure, the psychologist Miller raised the topic—but he immediately dropped it like a hot potato. He was, he said, "unclear" whether epistemic (cognitive) systems should be defined as animate or inanimate. The advantage of defining them as animate was that "we cut artificial intelligence free to develop in its own way, independent of the solutions that organic evolution happens to have produced" (Miller 1978: 9). (This remark predated the concept of strong A-Life by a decade: clearly, Miller thought it obvious that computers and life are incompatible.) But whether it made "any real difference" in conceptualizing the study of *mental* processes was "unclear".

Most analytic philosophers tacitly assumed some life-mind linkage—which would imply that if computers aren't alive then they aren't psychological systems. They evidently thought this point so obvious that, even when they bothered to state it explicitly, they didn't offer any arguments for it.

Scriven, for instance, confidently declared—without giving reasons—that "Life is itself a necessary condition of consciousness" and that "Robots ... are composed only of mechanical and electrical

parts, and cannot be alive" (Scriven 1953: 233). Lucas hinted at a similar position in his own reply to Turing's 1950 paper (see Section v.a). Geach insisted that AI systems can't have beliefs and intentions because they're "certainly not alive" (Geach 1980: 81). And some, such as Searle (1980, 1992) and Ruth Millikan (1984), explicitly linked intentionality with biology (neurochemistry and evolution, respectively). But even they didn't discuss the nature of life as such.

Two exceptions that proved (i.e. tested) the rule were Putnam's (1964) paper on "Robots: Machines or Artificially Created Life?" and Geoffrey Simons' (1983) book *Are Computers Alive?*

Despite its title, Putnam's paper focussed mainly not on life, but on consciousness. At one point, Putnam endorsed Ziff's claim that it's an "undoubted fact" that if a robot isn't alive then it can't be conscious. But he was relying on "the semantical rules of our language", not on any quasi-explanatory relationship between life and mind.

He also said (this time, disagreeing with Ziff) that something which is clearly a mechanism might be alive. Again, however, this was linguistic philosophy in action. Sometimes, Putnam heretically recommended *changes in meaning* due to new scientific data, as he did when countering Malcolm's account of dreaming (Putnam 1962). But in the paper on robots and life, he was talking only about what *current usage* allowed one to say (or imagine) without contradiction. The nearest he got to discussing a substantive claim about life was to scorn the suggestion that the primary difference between a robot and a living organism is the "softness" or "hardness" of the body parts (1964: 691).

Much later, Putnam's paper was discussed at length, and accused of incoherently combining Aristotelian and Cartesian views (Matthews 1977). At the time, however, it didn't prompt philosophical interest in the concept of life.

Simons, writing twenty years after Putnam, used concepts drawn from GOF AI and cybernetics to claim that computers can be *really* alive, and *really* intelligent. He specifically denied that the genesis of the system is relevant to whether it's alive: "A mechanically *assembled* [i.e. not evolved or self-constructed: see below] system may reasonably be regarded as living ...." (1983: 23). However, his argument was neither deep nor convincing, and (deservedly) attracted little attention.

### 3 b: Functionalist approaches to life

With the rise of A-Life in the late 1980s, the nature of life became an inevitable topic for computational research. Inevitable, but in practice not central: most A-Life workers focussed on other questions, maintaining a diplomatic silence on this one. Some of their colleagues, however, were more bold.

The relevant discussions were guided by two radically opposed philosophies. (Sounds familiar?—see Section vi.b.) These were functionalism and metabolic holism, a special case of dynamical systems theory.

Functionalism, in this context, is the view that the characteristics of life (see Chapter 15.ii.b) can be described by informational concepts. So self-organization involves the appearance of new levels of order, abstractly defined. Autonomy, emergence, development, adaptation, responsiveness, and evolution concern various types of structure, process, and control. Even reproduction (on this view) can be defined informationally, as self-copying.

The one exception is the concept of metabolism, which concerns not information but energy. Thoroughgoing A-Life functionalists weren't worried by this, as we'll see. But their opponents argued that they should be.

It's often assumed (wrongly) that all A-Life workers are thoroughgoing functionalists. This is largely because Christopher Langton, following John von Neumann's lead, wrote this position into his definition of the field in 1986 (Chapter 15.ii.b and ix).

Moreover, he drew the obvious implication: a licence for strong A-Life. If living self-organization is definable in logical terms, then a virtual "creature" implemented in computer memory that satisfied these abstract criteria—whatever they are—would be genuinely alive. ("Whatever they are", because definitions differed. For instance, Langton suggested including the lambda parameter, Andrew Wuensche the Z-parameter: Chapter 15.viii.a.)

Some A-Life colleagues were quick to join Langton in this claim. Thomas Ray, for instance, declared:

The intent of [my] work is to synthesize rather than simulate life .... To state such a goal leads to semantic problems, because life must be defined in a way that does not restrict it to carbon-based forms. It is unlikely that there could be general agreement on such a definition .... Therefore, I shall simply state my conception of life in its most general sense. I would consider a system to be living if it is self-replicating, and capable of open-ended evolution [generating] structures and processes that were not designed-in or preconceived by the creator (Ray 1992: 372).

As we saw in Chapter 15.vi.b, Ray's *Tierra* system did indeed generate phenomena not designed-in by Ray. These included co-evolving parasites, hyper-parasites, cheaters, and symbionts. Ray's response was a curious combination of modesty and hubris:

[The] results presented here are based on evolution of the first creature that I designed, written in the first instruction set that I designed. Comparison with the [virtual] creatures that have evolved shows that the one I designed is not a particularly clever one .... It would appear then that it is rather easy to create life (p. 393).

As for the problematic concept of metabolism, Ray said two things. On the one hand, the computer consumes physical energy too. On the other, the equivalent of metabolism can be functionally defined:

In studying the natural history of synthetic organisms, it is important to recognize that they have a distinct biology due to their non-organic nature. In order to fully appreciate their biology, one must understand the stuff of which they are made. To study the biology of creatures of the RNA world would require an understanding of organic chemistry and the properties of macro-molecules. To understand the biology of digital organisms requires a knowledge of the properties of machine instructions and machine language algorithms (p. 397).

I will discuss the inoculation of evolution by natural selection into the medium of the digital computer. This is not a physical/chemical medium; it is a logical/informational medium .... Evolution is then allowed to find the natural forms of living organisms in the artificial medium. These are not models of life, but independent instances of life (Ray 1994: 179).

For some broadly functionalist A-Life scientists, this was a step too far—and *much* too far for most philosophers (e.g. Harnad 1994; Olson 1997). Those A-Life colleagues were content to interpret most of the characteristics of life in informational terms—but not metabolism, which is irredeemably physical. However, since they defined metabolism as mere energy dependency, their rejection of Ray's position was intuitive rather than strongly argued (see below).

Many A-Life colleagues simply avoided the question, by way of the "diplomatic silence" mentioned above. They were interested in studying specific aspects of life, such as evolution or flocking, not in discussing its general nature. They were even less interested in considering the "strong A-Life" scenarios sketched by Langton and Ray—and later by Steve Grand (1958—).

Grand's first claim to fame was that he designed the hugely popular computer game *Creatures*. This swept the world in the early 1990s (see Chapter 13.vi.d), and was still being widely celebrated—for broadly counter-cultural reasons—in the new century (Kember 2003: 91-105).

*Creatures* enabled the user to evolve unusually sophisticated computer creatures (Grand and Cliff 1998). Their neural-network brains supported simple learning, and included 'neuromodulators' as well as several types of 'neurone.' The creatures also had a simulated biochemistry, with the potential to model a large number of metabolic and hormonal functions, from digestion to ovulation. As a piece of life-like software engineering, it was way beyond the general state of the art when it appeared, and is still impressive. Indeed, it could conceivably be used as a powerful testbed for AI models of motivation and emotion such as those discussed in Chapter 7.i.e-f (Boden 2000b).

Grand's current technical aim is to build an "imaginative" robot called Lucy, whose intelligence will emerge "naturally"—and holistically—from its 100,000-neurone hardware (Whitby and Grand 2001). As he points out, this attempt to build Dennett's (1978c) "whole iguana" is very different from MIT's Cog project, with which Dennett himself was involved (see 15.vii.a).

The Cog robot was carefully designed module by module, bits of its "intelligence" being successively bolted on. Grand, by contrast, wants an already integrated intelligence to emerge from a relatively unorganized base. Rather than providing Lucy's brain with spatial maps or orientation columns, for instance, he hopes that these would emerge spontaneously (much as ocular dominance columns arose in the work of Christoph von der Malsburg and Ralph Linsker: see Chapter 14.vi.b and ix.a). And the robot would learn to perform "voluntary" actions by associating the image (representation, model) of the desired action with the muscle movements required to achieve it (compare Marr's theory of the cerebellum: 14.v.c).

The Lucy project is startlingly ambitious—I'm tempted to say, utterly impracticable. But the A-Life expert David Cliff (p.c.) believed *Creatures* to be utterly impracticable too, when first consulted by the games company to whom Grand had offered it. Given what Grand had told them it could do, it must—so Cliff thought—be either hype and/or a superficial con-trick, carefully tailored to present a convincing 'demonstration.' (Even the impressive SHRDLU, you'll remember, could handle only the one conversation without tripping over its toes: 9.xi.b.) And the fact that it had been two-finger-typed on Grand's bedroom computer wasn't promising. Not until he got down into the machine code was Cliff convinced—at which point he suggested how it could be improved still further, using some of the ideas discussed in Chapters 7.i.f and 15.vi-ix.

Grand didn't know about those ideas already, because he's an autodidact. As such, he's undeterred by received academic opinion. And he's a highly creative computer engineer, who's already designed one apparently impossible system that does just what it was intended to do. He's thus in an entirely different class from the self-publicizing roboticists Kevin Warwick and Hugo de Garis, on whose 'research'—technical no less than philosophical—I forbear to comment, for fear of scorching the page.

I wouldn't bet a large sum of money on Lucy. And I don't agree with those cultural commentators who claim that Grand is "one of the 18 scientists most likely to revolutionise our lives in the coming century" (n.a. 2000b). Nevertheless, as Richard Dawkins has remarked, "If anybody can pull off a spectacular breakthrough, it'll probably be him" (Whitby and Grand 2001: 13). (For the most recent status-report on Grand's progress, see his website at <http://cyberlife-research.com/people/steve/>.)

At the turn of the century, Grand (2000, 2003) made a number of highly provocative claims about the philosophical significance of his own past and future work. He sees his virtual creatures as more than merely *life-like*: they are "sort of alive", or even "a sort of life". When challenged on this point, he insists (p.c.). Grand is an autodidact in philosophy too, but here there's no good reason to give him the benefit of the doubt. Whereas *Creatures* (considered as technology) clearly does what he said it would do, his philosophical arguments are challengeable—and, in my view, as mistaken as Ray's. Strong A-Life is no more plausible in *Creatures* than in *Tierra*—and even Grand's predicted robot Lucy wouldn't count as genuinely *alive* (see the discussion of metabolism, below).

Where the general public were concerned, Lucy made something of a splash. Although it must be said—and often is said, by other roboticists—that if Grand hadn't fitted a furry gorilla-face onto the head, and if he'd called it "Robot 37" instead of "Lucy", people wouldn't have been quite so interested. (Similarly, the young Minsky's robot arm aroused no attention until he put a shirtsleeve on it: see 1.iii.h.) Quite apart from the over-excitement of the journalists (the same old story!), a number of commentators have picked up on it as an expression of wider cultural concerns.

The anthropologist Lucy (sic!) Suchman, for example, who cast doubt on GOF AI planning some twenty years ago and focussed on *human-machine communication* soon after that (13.iii.b), described her robotic namesake as one among the disturbing category of the "almost human" (Castaneda and Suchman forthcoming; cf. Suchman 2004).

Besides the familiar anthropologists' fare of totems and other things "doing duty as persons", these include children, non-human primates, and AI/A-Life machines. The cultural status of children has been a focus of commentary at least since Jean-Jacques Rousseau (1712-1778), and twentieth-century developmental psychology has helped fuel this fire. As for primates, advances in field ethology have led to the culturally problematic Great Ape programme (7.vi.f). The eighteenth-century automata (2.i.b) challenged contemporary notions of the person (Riskin 2003). Now, as Suchman pointed out, various actual and imaginary AI projects are exciting comment not only in the philosophy of mind but in our wider culture too.

Lucy (which Suchman discusses at length) is only one example of the "almost human" produced by AI/A-Life. Cog, and especially its successor Kismet (see 13.vi.d), are others. The feminist philosopher Evelyn Fox Keller (forthcoming), for instance, sees some "serious anxieties" with respect to providing Kismet and the like with

facial expressions that reliably elicit emotional reactions in human viewers. (She's particularly worried by the plan to use robots like Brian Scasselatti's Nico to *test* theories in human developmental psychology.) Still other almost-humans—all media darlings in their day—include ELIZA, expert systems, AI agents ("softbots"), VR avatars, Turing's computer conversationalist, Stanley Kubrick's HAL, and Stephen Spielberg's David.

The behaviour—and man-machine interactions—of many of these systems is far more humanlike than Lucy's is. But because Grand, besides providing the superficial furry face, speaks of *life* as well as *mind*, his work aroused more outside comment than most. In addition, his A-Life system is not virtual/intellectual (as softbots are) but *embodied*—or at least, *material*. It's therefore of interest to those commentators, including phenomenologists and many feminist philosophers, for whom the downplaying of embodiment in the analytic/scientific tradition has been a fundamental mistake (Haraway 1997: 186, 302f. and passim; Kember 2003: 105-115, 198ff., and passim).

Suchman (like Haraway, and also Clark: vii.d, above) takes personhood, in whatever culture, to be constituted not by an individual person-in-the-mind but by the nexus of social relations and interactions available. On that view, the cultural status of robots, and other AI/A-Life systems, is determined less by their seeming intelligence than by the pattern of interactions we choose to engage in with them. But the influence is reciprocal: insofar as we do engage with them, we modify our own self-image in various subtle ways (cf. 13.vi.d).

#### 4 c: The philosophy of autopoiesis

Some A-Life researchers dismissed all these science-fictional scenarios *because they were fundamentally opposed to functionalism in the first place*. Among these were the proponents of Maturana's theory of "autopoiesis".

This was perhaps the best-developed philosophy of metabolic holism. (The competing candidate is the work of Pattee's group: see below.) It even inspired several computer models of biochemical autopoiesis (Zeleny 1977; Zeleny, Klir and Hufford 1989), and a wide range of work in A-Life (McMullin 2004). This included work in "wet" A-Life, in which biochemical autopoiesis *as such* was studied too (see Chapter 15.x.b). (Bachman et al. 1990; Walde et al. 1994).

Originated in the 1960s, Maturana's theory was strongly influenced by Heinz von Foerster's cybernetics. Despite the fact that an English translation was published over a quarter-century ago in the highly respected "Boston Studies in the Philosophy of Science" (Maturana and Varela 1972/1980), it has remained a minority taste. It's clear from my personal acquaintance that many philosophers have never even heard of it.

One reason, no doubt, is its rebarbative vocabulary and unrelenting abstraction. Also, it has some highly counterintuitive implications, as we'll see. Nevertheless, it offers a principled way of grounding mind in life. Rather than arguing (like Searle) that neuroprotein happens to cause intentionality, as chlorophyll happens to cause photosynthesis, this view grounds intentional categories in an essentially autopoietic biology.

For Maturana and Varela (1972/80), life is "autopoiesis in the physical space". Autopoiesis in general is defined as the continuous self production of an autonomous entity. As they put it (you're advised to take a deep breath here):

An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components that produces the components which: (i) through their interactions and transformations continuously regenerate the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in the space in which they (the components) exist by specifying the topological domain of its realization as such a network (Maturana and Varela 1972/80: 79).

Or more colloquially, an autopoietic system “pulls itself up by its own bootstraps and becomes distinct from its environment through its own dynamics, in such a way that both things are inseparable”. This type of self-organization can occur in the world of human communication, in which case we have some kind of social institution (cf. Teubner 1987, 1993). But when it happens in the physical world, we have a living organism.

The autopoiesis concerned here is a special case of homeostasis (see Chapter 4.v.c), where what’s preserved isn’t one feature, such as blood temperature, but the organization of the system as a unitary whole. This requires the self-creation of a unitary physical system, by the spontaneous formation of a boundary—at base, the cell membrane—and the continuous generation and maintenance of the body’s own components.

For Maturana and Varela, *body* and *embodiment* are autopoietic categories. So too are *cognition*, *communication*, *meaning*, and *language*, all of which they defined in terms of the interactions of living things. In the more accessible version of their theory that appeared around 1990 (Maturana and Varela 1987, 1992), and in Varela’s book coauthored with cognitive psychologists (Varela, Thompson and Rosch 1991), they focussed on human language, understanding, society, and consciousness—all described as necessarily rooted in our biology.

In fact, they were overly liberal with their ascriptions of intentionality (Boden 2001), for they declared that “Living systems are cognitive systems, and living as a process is a process of cognition” (1972/80: 13). Taken seriously, this extends knowledge even to algae and oak trees. One can—and should—express the idea that algae and acorns are pre-adapted to their environment without using the concept of *knowledge*. Such over-liberality was an occupational hazard for cyberneticians: as we saw in Chapter 4.v.e, Gregory Bateson had similarly attributed *knowledge* to redwood forests, and *mind* to whirlpools and oscillating electrical circuits.

From the autopoietic viewpoint, both strong A-Life and strong AI are absurdities. For computers aren’t autopoietic systems. Even self-assembling robots, if assembled from manufactured parts as opposed to being self-organized by some alien biochemistry, wouldn’t be alive. (Nor would they have bodies.) Consequently, robotic intelligence is impossible too.

Autopoietic theory is a special case of the general (anti-functional) position that metabolism is essential for life. Believers in strong A-Life (such as Ray, quoted above), when confronted with this view, typically pointed out that computers consume energy too. They sometimes added that the “physics and chemistry” of their virtual creatures is constituted by the computer’s memory and operating system.

A number of philosophers, some of whom weren’t committed to autopoietic theory, replied that metabolism is more than mere energy dependency. Rather, it’s the self production and self maintenance of

the physical body by energy budgeting, involving self-equilibrating energy exchanges of some *necessary* complexity (Pattee 1989; Cariani 1992; Sober 1992; Boden 1999). They argued that strong A-Life is possible only if virtual systems can metabolise *in the sense just given*, or if metabolism is inessential for life. But neither alternative is tenable.

Living ‘tin-can’ robots are also excluded by this approach. Only robots powered by complex biochemical cycles of synthesis and breakdown would be truly alive, and truly embodied. This is the basis of the intuition scorned by Putnam, that “softness” and “hardness” matter (see above).

Elliott Sober (1948–) cited other biological properties, besides metabolism, in arguing against strong A-Life (Sober 1992). Digestion and predation (for example) each relate an organism to something outside itself, where that “something” is essentially physical. Both can be realized in multiple ways (defined by biochemistry and behaviour), but in every instance some physical organism has to interact with—hunt, eat, transform—another. Like metabolism itself, these features can be usefully simulated by A-life models. But they can’t be replicated, so strong A-Life is impossible.

Sober’s argument would be endorsed by autopoietic theorists. But to see metabolism as essential for life isn’t necessarily to accept autopoietic philosophy. For this has some surprising implications, which many people reject. One was noted above, namely, the conflation of life and cognition. Another was remarked in Chapter 15.viii.b: the embargo on terms such as *input*, *output*, *function*, *feature detector*, and *representation*. Two more concern features often listed in definitions of life: reproduction and evolution.

Maturana and Varela’s claim that the formation of the cell membrane is *the* fundamental phenomenon of biology, and that life involves the “total subordination of [all the processes of change within] the system to the maintenance of its unity” (1972/80: 97), implied that reproduction isn’t essential for life. For them, this process is not (as functionalists claim) informational self-copying, but the formation of new autopoietic unities from previous ones. It follows that life is prior to reproduction (pp. 105–107). This wasn’t a merely conceptual point, but a substantive biological hypothesis: that the earliest living organisms needn’t have been able to reproduce (Boden 2000b).

Evolution, also, was seen by them as inessential, because it requires reproduction. (Inessential for life, but not for what’s normally regarded as knowledge: they admitted that only evolution can generate the complex organisms typically credited with cognition.)

This conclusion, though unusual, is less controversial. For, *pace* Ray, and many theoretical biologists too (e.g. Maynard-Smith 1996), there are three independent arguments against defining life in terms of evolution. First, populations, not individual organisms, would be paradigm cases of life. Second, creationism would be conceptually incoherent, not just false. And third, a population in evolutionary equilibrium wouldn’t count as alive.

## 5 d: Evolution, life, and mind

Some philosophers of A-Life, nevertheless, took evolution (together with metabolism) to be the sort of self-organization which characterizes life. Pattee was an early example, followed by his students Rosen (1985, 1991) and Cariani (1992, 1997). He’d modelled co-evolution in the 1960s (see Chapter 15.vi.a). Subsequently, he focussed on the emergence of new phenotypic structures and functions.

A crucial example, for Pattee (1985), was novel types of “measurement”, or classification. These were understood as ranging from enzyme activity to sensory perception—as in the evolution of new sensory organs (see Chapters 4.v.e and 15.vi.d). Pattee’s concept of measurement was intriguingly similar to Smith’s “participatory registration”—but, unlike Smith, he retained the first definition of computation distinguished in Section ix.a. So he specifically dismissed strong A-Life, arguing that measurement requires physical interaction, which can’t be realized by formal computational systems. *A fortiori*, no novel biological functions can emerge in formal evolutionary systems (15.vi.d). He did allow, however, that “weak” A-Life modelling (simulation) could help clarify central biological and psychological concepts.

In the 1990s, another philosopher of A-Life argued that evolution is an essential criterion. Mark Bedau (1954–) explicitly accepted the three counterintuitive implications mentioned above, because of the explanatory power gained by defining life in evolutionary terms (Bedau 1996). And this explanatory potential, he said, was augmented by A-Life. In presenting his account of “supple adaptation” (alias evolution), he argued that A-Life modelling can deepen our understanding of life as such, because it helps us to study evolution in dynamic and quantitative terms. Moreover, he extended his evolutionary argument from life to mind (Bedau 1999, in prepn.).

A-Life philosophers weren’t the only ones to link life and mind. Others, too, had grounded knowledge and meaning in biological evolution. Dennett had sketched an evolutionary account of meaning in *Content and Consciousness*, although philosophers then were more interested in other aspects of his work (see Section iv.a). By the mid 1980s, however, two influential examples of teleological or evolutionary semantics had appeared.

The philosopher of science David Papineau (1947–) argued that the content of beliefs depends on how they guide actions to satisfy desires, whose content is basically determined by natural selection (Papineau 1984, 1987). Similarly, Millikan (1933–) grounded intentionality in evolutionary history (Millikan 1984). Her book title was deliberately provocative: *Language, Thought, and Other Biological [sic] Categories*. This was guaranteed to raise philosophical hackles in devotees of the later Wittgenstein, and neo-Kantians in general (see Sections vi–viii, above).

Millikan upset many science-inclined naturalists too, by giving more philosophical weight to evolution than to neuroscience. Thus she argued that a perfect simulacrum of a human being, magically constituted in the middle of a swamp by a sudden combination of the relevant molecules, would have *no* beliefs, desires, or other intentional properties (1984: 93, 337f; 1996; cf. Boorse 1976). It would, of course, utter the very same words as a human being would, if engaged in ‘conversation.’ For all the language-relevant events in the swamp-man’s brain (and ears, and lips ...) are, by hypothesis, identical with those of a person. But it wouldn’t be a genuine conversation—for, on the swamp-man’s side, no meanings or intentions would be being expressed. (In her defence, one could point out that we accept thermodynamics *even though* it allows the theoretical possibility of a snowball in Hell: is swamp-man any more implausible?)

This imaginary example highlighted her central—and controversial—claim, that current meanings depend in part on events that happened millions of years ago. Millikan was saying, in effect, that Searle had been wrong about the “something more” that’s needed *in principle* for intentionality. According to her, it’s not neurochemistry as such that grounds meaning—nor even

neurochemistry in interaction with the body and environment. Only evolutionary history can fix the system’s semantics.

If Millikan’s (or Papineau’s) version of biological semantics is correct, then no ‘ready-made’ AI-system, nor even a self-organizing—but non-evolutionary—A-Life system, could enjoy mind, intelligence, or meaning.

However, evolutionary semantics was later related to research in evolutionary robotics (Boden 2001). We saw in Chapter 15.vi.c that a robot’s neural-network ‘brain’ may evolve ‘feature detectors’ analogous to those found in mammalian visual cortex. So a mini-network may evolve that’s sensitive to a light-dark gradient at an orientation matching one side of a white cardboard triangle, and that’s used by the robots as a navigation aid (Harvey, Husbands and Cliff 1994; Husbands, Harvey and Cliff 1995). Such examples challenge Searle’s (1980) view that the “meaning” of a computer model must always be derivative, and arbitrary to boot (see Section v.c).

One could debate whether the feature detector means “light-dark gradient sloping up and to the right” as opposed to “left side of the white triangle”. But similar difficulties attend the ascription of non-conceptual content to animals. (Are bug detectors really *bug* detectors, whether for the frog or for the frog’s brain?—see Chapter 12.x.f and Cussins 1990: 416f.).

The important point is that the various meanings one might want to ascribe to the robot aren’t arbitrary. Nor are they derivative, based only in the human purposes involved in their design. They aren’t based purely on causal regularities, either. They spring to mind as candidate meanings because the mini-networks concerned have evolved, within that task environment, to discriminate certain visual features and guide the robot’s movements accordingly. That is, they’re environmentally, enactively, and evolutionarily grounded.

However, to say these A-Life “meanings” aren’t arbitrary isn’t to say they’re genuine. There’s no consensus among A-Life researchers on whether evolutionary robotics could produce real intentionality. For the pure A-Life functionalist, it could: the triangle detector is a primitive case, and more advanced (animal-like) examples would embody richer meanings. For Maturana, it couldn’t: evolution and intentionality can occur only in biological organisms—so quasi-evolved robots can quasi-embody only quasi-meanings.

Nor is there a consensus among philosophers unconnected with A-Life, for the nature of life, mind, and the life-mind relation remain controversial.

Not everyone accepts an evolutionary semantics, for example. A causal semantics can’t support the commonsense intuition that mind can arise *only* from life, unless the relevant causal relations can be shown to arise *only* in living things. And a model-theoretic semantics can’t support it at all.

The competing A-Life methodologies of the early 1990s were systematically compared, and related to earlier philosophies of life, by Peter Godfrey-Smith (1994). He distinguished three dimensions of variation: internalism and externalism; asymmetrical and symmetrical externalism; and weak and strong versions of the continuity of life and mind.

Internalist approaches see life as autonomous self-organization, wherein internal constraints govern the history and interactions of the constituent units of the system. Examples include autopoietic theory and Stuart Kauffman’s autocatalytic networks (15.viii.b). Externalist approaches explain the system’s internal structure primarily as a result of its adaptive interactions with the environment. Work on evo-

lutionary robotics is one example.

The asymmetric externalist emphasizes the organism's adaptive responses to its environment. By contrast, the symmetric externalist pays attention also to the active role of the adaptive organism in shaping that environment. Examples are situated robotics, and Ray's or Pattee's models of co-evolution, respectively.

Finally, the weak continuity theorist sees mind as emerging only from life, but as significantly different from it, whereas the strong continuity theorist regards mind and life as ontologically similar, sharing basic organizational principles. Descartes wasn't a continuity theorist at all, for he saw mind and living bodies as utterly distinct (see Chapter 2.iii and Matthews 1977). Examples of strong continuity theorists include the *Naturphilosophien* (Chapter 2.vi), the cybernetics movement (Chapter 4.v-vii), and autopoietic theorists. Arguably, they also include philosophers of non-conceptual content (Chapter 12.x.f) and participatory computation (see Section ix.e, above). And someone who argues that not all living things are cognitive systems (see above), is supporting weak continuity in that respect.

However, "mind" covers a number of abilities, and some of these may be strongly continuous with life whereas others aren't. Language has often been seen as a cut-off point. For instance, we saw in Chapter 2.ii.a.g that Aristotle was a strong continuity theorist for perception and autonomous movement, but perhaps not for human reason (cf. Matthews 1992). Heideggerians who confine *dasein* to human beings, or Wittgensteinians who ascribe intentionality only to linguistic concepts, count thus far as weak continuity theorists. But some neo-phenomenologists (such as Clark and Wheeler) ascribe intentionality to non-human animals, too.

Analogously, many AI connectionists allow that GOF AI insights will be needed to model the 'logical' aspects of human thinking (see Chapter 12.viii-ix), whereas some dynamical theorists deny this (Section vii.c, above). And *nouvel AI* (a label recalling the minimalism of *nouvel cuisine*) insists that AI must be grounded in 'lower' abilities, like those of our evolutionary precursors, whether or not it has to add GOF AI methods on top.

In sum, the relation between life and mind is still highly problematic. That applies to work in AI/A-Life, and to philosophy too. The commonsense view is that the one (*life*) is a precondition of the other (*mind*). But there's no generally accepted way of proving that to be so.

# CHAPTER 17: WHAT NEXT?<sup>1</sup>

Margaret A. Boden<sup>2</sup>

## Abstract.

This is a re-formatted version of Chapter 17 of M. A. Boden, *Mind As Machine: A history of Cognitive Science (Vols 1–2)*, Oxford University Press, Oxford, 2006. The chapter title is *WHAT NEXT?* which is also the topic of the final discussion session of the AIIB symposium, which Professor Boden has agreed to introduce.

The book has a large bibliography. All the references in this extract are to items listed in the bibliography. The section numbers correspond to sections in the book.

It is included on the AIIB web site with the kind permission of the author.

## 1 Introduction

This chapter might have joined Chapter XI of *Through the Looking Glass* as one of the shortest in the English language. For in response to “What next?”, what is there to say but “Who knows!”? Fundamental advances, in particular, are unforeseeable. As Captain Cook’s biographer put it, “Genius, of whatever sort, takes us unawares: is not, even in retrospect, deducible” (Beaglehole 1974: 3).

One doesn’t have to adopt a Romanticist view of creativity, nor a literal interpretation of “genius”, to agree with that. Creative ideas are unpredictable for a number of very different reasons, not all of which will be mentioned here (but see Boden 1990a: chap. 9). And some are more unpredictable than others. Even a carefully designed technological artefact will have some unpredictable features, as we saw in Chapter 8.v.b (Ziman 2000b). A relatively speculative creative idea can be more surprising still. (Its social impact is even less predictable: Tim Berners-Lee himself couldn’t have foreseen that in a mere three years, from 1993 to 1996, the Web would grow from 130 sites to over 600,000—Battelle 2005: 40.)

More precisely, *historically* creative ideas, never generated by anyone before, are unforeseeable. Creative ideas that are new only for the person concerned can sometimes be foreseen, and even deliberately brought about, by other people—think of Socratic dialogue, for example. In what follows, I’ll use “creative” to mean historically creative.

## 17.i: What’s Unpredictable?’

One source of unpredictability is serendipity: the finding of something valuable without its being specifically sought. Since this is unexpected by definition, prediction simply isn’t on the cards when it’s involved. The classic case in science is Alexander Fleming’s noticing the dirty dish of agar-jelly, which led eventually to the discovery of

penicillin. In cognitive science, the part-accidental discovery of several visual feature detectors is another illustration (Chapter 14.iv.a-b). In both cases, of course, a good deal of careful and systematic work had to follow the initial observation, before anything worth calling a “discovery” could be achieved.

Another source of unpredictability is change in the wider cultural context. The generation—and still more, the acceptance—of new scientific ideas can be discouraged, encouraged, and even part-guided by social/political factors (see 1.iii.b-d and 2.ii.b-c). The positive reception of heterarchy (10.iv.a) and distributed cognition (13.iii.d-e), for instance, was influenced by political ideology, and of expert systems by nationalism and economics (11.v). Even if cultural changes could be predicted (which they can’t), their effects on contemporary scientific thinking could not.

This applies also to shifts in the power relations between the various groups/disciplines within cognitive science, and so to *what will be seen as cognitive science* in the future:

Some 25 years after its various beginnings there still is no such thing as a core cognitive science. Depending on where one looks, which departments one queries, who one’s friends are, the core of cognitive science will be asserted to be neurophysiology, psychology, artificial intelligence, linguistics, or some more vague concept like human/machine interaction or symbolic or connectionist modeling. The result may not have been cognitive science, but it has been exciting and scientifically fruitful. It has created a community of interests and increased interdisciplinary communication. But as of now there are still viable independent cognitive sciences such as neurophysiology, linguistics, and psychology that flourish with or without the cognitive science label or affiliation. *It is difficult to say at this point where this will lead* (Mandler 1996: 23; italics added).

A specially important type of cultural change concerns new technology. Sometimes, the new instruments are needed in order to do things that obviously needed doing: for instance, developing micro-electrodes to record from the cell-body of a single neurone—or anyway, a very small number thereof (2.viii.e). Such cases are relatively predictable: it was clear that people would try, and probable that someone would eventually succeed. Other technological tools may come as more of a surprise, at least to people in other areas of science: a few biophysicists may have been able to predict brain-scanning techniques, but psychologists couldn’t.

Technology can be used to prompt new concepts, as well as to find new data. Indeed, technical ideas have been transmuted into psychological theories on many occasions (Gigerenzer 1991b, 1994). Moreover, machines have been used as analogies for the brain for hundreds

<sup>1</sup> ©Margaret A. Boden 2006

<sup>2</sup> University of Sussex, UK, M.A.Boden@sussex.ac.uk

of years (Fryer 1978)—juke-boxes included (2.viii.f). The latest, of course, is the computer itself. So sceptics often say that it's just the latest in a long line of such analogies, to be displaced eventually by some unforeseen invention coming who-knows-when.

Well, yes and no. In cognitive science, the computer isn't merely a superficial analogy, a metaphor fished out of the memory—perhaps for purposes of popularization—after the real scientific work has been done. On the contrary, it provides substantive concepts in psychological and neuroscientific theories.

The computational concepts concerned were diverse even at the outset (4.ix), and have multiplied over the years (16.ix). Besides symbolic, connectionist, and evolutionary AI, they include dynamical systems described by differential equations (14.vi, 15.viii-ix and xi). The future may well hold unpredictable new machines, even less imaginable today than quantum computers are. But cognitive scientists believe that only *some sort of computational machine* will be relevant. For their key claim is that mind can be explained (not by today's ideas about computation, but) by *whatever theory turns out to be the best account of what computers do* (Chrisley 1999; see 16.ix.f). In that sense, they would endorse Philip Johnson-Laird's remark that: "The computer is the last metaphor; it need never be supplanted" (1983: 10).

Other psychological reasons for unpredictability apply to all instances of creativity. They include the rich idiosyncrasy of human minds and the relative—though only *relative*—freedom of creative thinking (13.iv; cf. 7.i.g). Introspectively, it may seem as though almost anything can happen—at least, according to the molecular biologist Francois Jacob:

Day science employs reasoning that meshes like gears .... One admires its majestic arrangement as that of a da Vinci painting or a Bach fugue. One walks about it as in a French formal garden ... Night science, on the other hand, wanders blindly. It hesitates, stumbles, falls back, sweats, wakes with a start. Doubting everything ... It is a workshop of the possible ... where thought proceeds along sensuous paths, tortuous streets, most often blind alleys (Jacob 1988: 296).

If the creative scientist himself "wanders blindly" much of the time, so much less can his thoughts be foreseen by other individuals—who don't even know his *present* thinking in much detail.

Moreover, some new ideas strike us as paradoxical, not to say crazy, even *after* they've occurred. That often happens in instances of "transformational" creativity, in which one or more dimensions of the previously-accepted style of thinking is/are radically altered or dropped (Boden 1990a: chaps. 3-4). The more basic the dimension, the more fundamental the conceptual change will be. In such cases, it's hard for the new idea to be understood, and even harder than usual for it to gain acceptance. *A fortiori* it's harder to predict.

In particle physics, that's par for the course. Freeman Dyson reported an encounter between Niels Bohr and Wolfgang Pauli, who'd given a lecture on his new theory:

Bohr rose to speak. "We are all agreed", he said to Pauli, "that your theory is crazy. The question which divides us is whether it is crazy enough. My own feeling is that it is not crazy enough" (Dyson 1958: 74).

And Dyson commented:

The objection that they are not crazy enough applies to all the attempts which have so far been launched at a radically new theory of elementary particles. It applies equally to crackpots. Most of the crackpot papers which are submitted to *The Physical Review* are rejected, not because it is impossible to understand them, but because it is possible. Those which are impossible to understand are usually published. When the great innovation appears, it will almost certainly be in a muddled, incomplete and confusing form. To the discoverer himself it will only be half-understood; for everybody else it will be a mystery. *For any speculation that does not at first glance look crazy, there is no hope* (italics added).

Cognitive science as a whole is less *rococo*, less conceptually bizarre, than particle physics. But several seemingly crazy ideas have found a respected place within it, after being fiercely resisted as "absurd"—perceptual expectancies, for example (Chapters 6.ii and 16.v.f), and object-oriented programming (10.v.d and 13.v.d). And remember the punchline of the quip about *Plans and the Structure of Behavior*: "... and Pribram believed it!" (6.iv.c). When people said that, they weren't dismissing those new ideas as worthless. They were allowing that they were weird-but-interesting, so worth thinking about. (Karl Pribram was made the fall guy because he'd recently defended a holographic theory of memory—hardly the usual bread-and-butter fare: 12.v.c. Nor is Bergsonian philosophy, but Pribram later dallied with that as well: 16.x.a. His reputation as a maverick was deserved. However, even those who called him "crazy Karl Pribram" later admitted that his strange ideas had "caught on," and that "his neurophysiological speculations are decades beyond other physiological work"—Walter Weimer, interview in Baars 1986: 309f.)

Many future contributions, too, will seem weird initially—though just how weird they'll need to be remains to be seen. The "particle physics" of the field is conscious experience. This has already prompted many highly counterintuitive theories, including some crackpot publications. I argued in Chapter 14.xi.e that a currently undreamt-of (i.e. crazy) approach will be needed to explain it.

Close runners-up in order of difficulty, and so in licensed craziness, are intentionality and computation. We saw in Chapter 16.ix.e how an extraordinary (crazy?) theory of those-two-together has come from an AI scientist/philosopher who thinks that "For sheer ambition, physics does not hold a candle to computer or cognitive ... science" (Smith 2002: 53).

Sometimes, experts declare future progress to be not so much unpredictable as impossible. This view was implicit in Thomas Watson's notorious remark in 1943, as IBM chairman, that "I think there is a world market for maybe five computers." (He died in 1956, so never knew just how wrong he was. But he wasn't alone: Howard Aiken, of all people, said "there will never be enough problems, enough work, for more than one or two of these computers"—Edwards 1996: 66.) And it was explicit in the advice given to Konrad Zuse in 1937 by Kurt Pannke, a manufacturer of specialized calculators:

"Someone informed me", Dr. Pannke began, "that you have invented a computing machine. Now, I don't want to discourage you from continuing to work as an inventor and from developing new ideas, but I must go ahead and tell you one thing: in the field of computing machines, practically everything has been researched and perfected to the last detail. There's hardly anything left to invent ...." (Zuse 1993: 42).

(To be fair to Pannke, he later changed his mind. He provided money to fund Zuse's home-based research, and recommended his machine to the German military—fortunately, with no effect: see 11.i.a.)

## 17.ii: What's Predictable?

I can't imagine anyone suggesting that there's "hardly anything left" to be discovered in cognitive science. But I've just allowed that creative ideas can't be predicted, only awaited. So perhaps I should now present you with an empty page, and leave it at that? After all, that's a respected rhetorical device. Laurence Sterne did it 250 years ago, when he declined to describe a beautiful woman in *The Life and Opinions of Tristram Shandy*, leaving it to the reader's imagination instead.

I don't have the courage to follow Sterne's example. But it wouldn't be appropriate in any case. For there is something that can be said.

All scientific research, in whatever domain, is located within some identifiable conceptual space where further creative exploration (and transformation) is clearly possible, and where some dimensions seem especially rich in potential with respect to current unsolved problems (Boden 1990a, 2004). Peer-review, especially of proposals for future research, depends on that fact. We can't predict the detailed outcome of such explorations and transformations, much as David Livingstone couldn't foresee his discovery of the Victoria Falls. But we can reasonably expect that if we follow *these* dimensions of the space (compare: the Zambesi river, the mountains glimpsed ahead ...), we'll find something of interest. That is, we can have intellectually defensible, if not infallible, hunches about where future discoveries are likely to occur.

Insofar as such predictions are possible, I've indicated mine already. The previous chapters have told "the story so far"—but always with an eye to possible future episodes. So the relatively small volume of *recent* work that I've mentioned was chosen not just because it's recent, nor even because it's intriguing. It was selected because I think it's promising, capable of development in ways that seem likely to be fruitful.

One way of justifying our hunches about where interesting new ideas are likely to arise is to rely on sub-hunches about *how* those ideas might be generated. In other words, some specific exploratory pathways are recognizable as familiar ones, because they've often been found to be fruitful.

- For instance, once a simple deterministic space has been defined, it's very likely that people will eventually try to complexify it in certain ways. So when John von Neumann defined the basic cellular automaton, he knew very well that probabilistic and even evolutionary CAs would be explored later (Chapter 15.v-vi).
- Again, once problem solving had been seen as a simple hierarchy (6.iii), it was inevitable that more complex and/or 'open-execution' plan hierarchies would be explored (10.iii.c). It was even a good bet that theories of problem solving would eventually be transformed by hierarchy's being made less pure (10.iv.a), or perhaps deliberately dropped (13.iii.b, 15.viii.a).
- Third, when Alan Turing wrote his morphology paper, he knew that increasingly complex systems of diffusion-reaction equations would be explored, once computer power allowed (15.iv).

(Similar remarks apply to creativity in artistic contexts. So, for example, it was nigh-inevitable that post-Renaissance composers would progressively complexify tonal harmony. And it was always on the cards that someone—it happened to be Arnold Schoenberg—would eventually transform the space of tonal music by dropping the home-key constraint altogether: Rosen 1976; Boden 1990a: chap. 4.)

In short, the common notion that creative thought is unpredictable because it's chaotic (in the everyday sense) is mistaken. There's significant method in creative madness. It's our tacit recognition of this fact which enables us to identify certain work as promising, even though we can't spell out the promises.

## 17.iii: What's Promising?

The recent empirical research that I see as promising in these terms includes the following (listed here in no particular order):

- hybrid systems I: symbolic/connectionist (Chapters 12.iii.d and ix.b, 15.viii.a)
- hybrid systems II: situated/deliberative (7.iv.b and 13.iii.c)
- hierarchical networks (12.viii.b and ix.b)
- connectionist work on the role of imagery-of-words (12.ix.e)
- statistical approaches to NLP (9.x.preamble and 9.x.f)
- integration of connectionist learning with detailed neurophysiological data (14.ii.d)
- modular and/or time-based neural networks (12.ix.a, 14.ix.g)
- programmed/evolutionary neuromodulation (14.ix.f)
- AI-evolved organic-silicon computing networks (14.ix.f)
- computational neuroethology (14.vii, 15.vii)
- insect navigation strategies (15.viii.a)
- types of cerebral representation, especially emulators (14.viii)
- the epigenesis of thought and language in normal and brain-damaged children (7.vi.g-i)
- models of clinical apraxia and aphasia (12.ix.b, 14.x.b)
- theories of control in hypnosis (7.i.h)
- brain-scanning, *provided that* it's related to specific psychological theories (14.x.b)
- developmental trajectories (12.viii.c-e and x.e)
- fast/simple heuristics (7.iv.f-g)
- the origin of specific bounds on human rationality (7.iv.h)
- cognitive technology, including virtual reality (10.i.h, 13.vi, 16.vii.d)
- computational theories of creativity (9.iv.f, 13.iv)
- evolutionary modelling (14.ix.d and f, 15.vi)
- achieving open-ended evolution and/or creativity (13.iv.c, 15.vi.d)
- mathematical analyses of dynamical systems (14.vi and ix.b, 15.ix.b and xi.b)
- homeostasis in CTRNs (15.xi.a)
- distributed cognition and agents (8.iii, 12.ii-vi and x, 13.iii.c, 15.viii-ix)
- computational architectures integrating knowledge, motivation, and emotion (7.i.e-g and 7.iv.b-c).

If forced to choose only one of these items, I'd pick the last: work on integrated mental architectures. Indeed, I did that on the 50th anniversary of the 1953 discovery of the double helix, when the British Association invited several people to write 200 words for their magazine *Science and Public Affairs* on "what discovery/advance/development in their field they think we'll be celebrating in 50 years' time". This choice reflected my own

longstanding interests in personality and psychopathology (Preface.ii). But it wasn't idiosyncratic: two years later, the UK's computing community voted for "The Architecture of Brain and Mind" as one of the seven "Grand Challenges" for the future (<http://www.uk.crc.org.uk/Grand.Challenges/index.cfm>). One member of the five-man committee carrying this project forward is Aaron Sloman, who's been thinking about architectural issues since the 1970s (7.i.f, 10.iv.b, and 16.ix.c). If progress is to be made on this front, my hunch is that his team will be in a good position to make it.

The Grand Challenges grew out of the UK government's "Foresight" Programme (instituted in 2003 for a ten-to-twenty year planning horizon), and in particular out of its "Cognitive Systems" Project. Naturally, government ministers aren't falling over themselves to help solve the problems of cognitive science for their own sake. For them, applications are all—whether in health, education, business, transport, arts and entertainment, or (of course) the military. But as the Project's official Report (n.a. 2004) makes clear, scientific and technological motives are often very closely related (and can be satisfied only by interdisciplinary thinking). It should be no surprise, then, that architectures to support "emotional" robots and "social" human-computer interactions are now being investigated at the behest of Whitehall—and, naturally, of the Pentagon too.

The strength and range of the list of "promises" given above show that cognitive science is still a fruitful "scientific research programme" (Lakatos 1970). *Mind-as-machine*, in both its incarnations (1.ii.a), has generated many suggestive theories. These have been amended—and sometimes dropped—on the basis of further advances in our understanding, but in many cases the central insights remain. Marr's work on vision is one obvious example (7.v.b-f), but others have been described in previous chapters.

The field's potential won't be unlocked without new psychological-computational *theories*. Greater computer power may well be necessary, but it won't be sufficient. Even quantum computers and hypercomputers won't suffice to fill the bill (16.ix.a). Fundamental scientific advance will need more Ideas, not just more Bytes. Likewise, more and/or fancier PET/fMRI brain scanning won't suffice either, even though it will often be useful (14.ii.d and ix.c).

On a higher plane of abstraction, I've discussed some recent philosophical research concerning:

- the nature of computation (16.ix)
- the variety of virtual machines (16.ix.a)
- conscious experience (14.x-xi)
- the nature of intentionality (16.x.d)
- the origin of conceptual content (12.ix.e and x.f)
- the nature/existence of non-conceptual content (12.x.f and 16.viii.b)
- mind and/as embodiment (16.vii)
- the boundary between self and world (16.vii.d)
- the nature of life, and its relation to mind (15.i and 16.x)
- the resolution/reintegration of neo-Kantian and analytic philosophical viewpoints (16.vii.b-d, ix.d-f, and x.a).

All of these matters will be key foci of effort and controversy in the foreseeable future. Indeed, they're so difficult, and so deep, that I expect them to remain key foci well over a hundred years from now. For as remarked at the outset of Chapter 16, philosophical problems don't get solved in a hurry.

Nor, in these cases, will they get solved in disciplinary isolation. They'll require fundamental *and reciprocal* advances in up to five fields: philosophy, psychology, anthropology, neuroscience, and AI. (Theoretical linguistics, as opposed to the philosophy of language, is less relevant here—unless we include *cognitive* linguistics: see 7.ii.preamble, 9.ix.g.)

## 17.iv: What About Those Manifesto Promises?

In Chapter 6.iv.c I said that a good way of judging how far cognitive science has succeeded is to compare it with the hopes/promises expressed in *Plans and the Structure of Behavior* (Miller, Galanter and Pribram 1960). By the turn of the millennium, virtually all of MGP's promises had been at least partially met. The "satellite images", and the Newell Test, outlined in Chapter 7.vii surveyed many different examples.

To mention just two:

- hypnosis has been demystified (along with multiple personality and religious experience): (7.i.h-i, 8.vi.b, and 14.x.c), and
- MGP's distinction between Plans as animal instincts and as human purposes is now far better understood (7.i and iv, 14.vi.c, and 15.vii).

Although discussions of these matters have been hugely complicated since they wrote their manifesto, today's answers are broadly consistent with theirs. For TOTE units were—deliberately—defined so abstractly that they covered *both* the inbuilt sensori-motor skills of crickets and hoverflies *and* the deliberative (and hypnotic) planning of human beings.

Neither "demystified" nor "far better understood" implies that all the relevant questions have been answered. Far from it. But we're much clearer now about just how MGP's questions can be profitably put.

Consider, for example, their nature-nurture distinction mentioned above. This simplistic duality has given way—within cognitive science, if not yet in the minds of the general public—to an epigenetic view of development. This view was already waiting in the wings before cognitive science got started (5.ii.c). Now, it's prominent in disciplines as varied as psychology, neuroscience, philosophy, A-Life, and robotics (7.vi, 14.vii and ix.c, and 15.viii.a).

There's no reason why this process should cease now. And it doesn't require every psychological question to be answerable by a simulation. For MGP's futuristic remarks concerned a general approach to the mind: computational theorizing, not necessarily computer simulation as such. We'll surely see many new computer models (some of which will reflect new findings in neuroscience). We'll probably see radically new *types* of model (16.ix). Functioning computer models can test a theory's implications and coherence more rigorously than any other method (7.iii.c). But the novel theoretical concepts are what's important, in understanding what sort of system, or machine, the mind/brain is.

One thing is beyond dispute: that the rich subtlety of human minds is even more awe-inspiring than the arch-humanist Wilhelm von Humboldt (9.iv) believed it to be. Indeed, I've already identified this realization as the major result of computational psychology as a whole (7.vii.a). It follows that it will never be possible to capture

every psychological detail, whether in a theory or a simulation. Predicting, explaining, or interpreting the specific thoughts/actions of individual people will always be largely “idiographic” (7.iii.preamble), a matter for the unargued intuitions of psychologists *qua* human beings, not for their deliberate conclusions *qua* scientists.

However, that doesn’t spell disappointment for MGP. For on the one hand, idiographic insights can often be enriched, and sharpened, by considering general mechanisms. Remember, for instance, the varied ways of expressing different types of anxiety in speech (7.ii.c). On the other hand, the prediction/explanation of highly particular personal matters wasn’t what MGP were aiming for. (Nor is this the aim of scientific psychology in general: 7.iii.d.) Rather, they hoped to understand how such phenomena are *possible*.

It’s not only MGP’s questions which can now be posed more fruitfully. The familiar puzzles that opened this story (1.i.a), many of them centuries old, have all been illuminated—and some even solved—by the successors of the visionary manifesto.

More answers will doubtless be found: the future of cognitive science will be as exciting as its past. But to say *what they’ll be* would be like an eighteenth-century Admiralty Board foreseeing James Cook’s extraordinary achievements in navigation and map-making: impossible.

# Forward and Inverse Models in Motor Control and Cognitive Control

Richard P. Cooper<sup>1</sup>

**Abstract.** It is now commonly accepted that the motor system makes use of so-called forward and inverse models in order to control the musculoskeletal system during rapid, skilled, motor behaviour. Inverse models are held to allow the system to determine the motor commands necessary to achieve a desired state, while forward models are held to allow the system to predict the expected sensory feedback of a motor command, allowing rapid error detection when actual and predicted feedback do not match. It has recently been suggested that these ideas from control theory might also be applied to the control of cognitive processes, allowing (for example) the cognitive system to anticipate processing conflict and pre-emptively minimise it by adjusting processing strategies or the allocation of processing resources. This paper reviews theories of cognitive control that are broadly consistent with the use of complementary forward and inverse models. It is argued that there is indeed a role for such models in cognitive control – particularly in relation to a putative monitoring function – but that the models involved are likely to be somewhat impoverished.

## 1 INTRODUCTION: MULTIPLE MODELS IN MOTOR CONTROL

Consider a skilled musician, say a pianist, sight-reading a piece of music. The cognitive task of translating the sensory input, for example an F-sharp followed by a lower C, into a sequence of motor commands – press the F-sharp key and then the C key to the left of it – is clearly a complex one. But consider the motor task. The motor system can't just execute two simple finger presses over the appropriate keys. It must work out which digit to use given the current position of the hand and digits. It may need to raise a digit, shift the position of the hand, or move a digit horizontally as well as vertically. In skilled sight-reading, all this must be done at pace – yet there is an inherent time lag in the transmission of neural signals. The motor system must be able to take account of such lag, but it can also not afford to wait for proprioceptive feedback from one motor command (such that it can access a representation of the location of all relevant skeletal components and the state of all relevant muscular components) before formulating the next. This, and related problems in motor control [1], have led researchers in the area of motor control to propose, following simple control theory, that the motor system makes use of so-called forward and inverse models in planning and regulating motor behaviour.

In general, a forward model is a representation of the future state of a system. The motor system might profitably make use

of such a model to predict its state following the performance of a motor command. That is, given the current state of the motor system and the intended action, a forward model allows the system to predict its future state. Wolpert and colleagues refer to the relevant type of model as a *forward dynamic model* [2].

For speeded motor activity, the prediction may be used in two ways. First, it allows the system to formulate the next motor command without waiting for proprioceptive feedback from the current command. How though does the motor system formulate this motor command? This is where an *inverse model* is held to play a role. Inverse models “invert the causal flow” [1]. Thus, given a future desired state, they generate the motor command that is required to bring about that state.

The second use of the predicted state is that it may be feed into a second forward model – a so-called *forward sensory model* [2] – that predicts the expected proprioceptive feedback based on the anticipated state. Any mis-match between this and the feedback subsequently received may be used by the motor system to fine-tune motor control in real time, as is required in skilled, speeded motor activity.

Wolpert and colleagues have amassed a substantial body of behavioural evidence that supports the claim that the motor system relies on internal forward and inverse models (see, e.g., [1,2,3]). These studies generally involve mapping moment-by-moment control of speeded movements in simple motor tasks, often in situations where proprioceptive feedback is manipulated so as to violate expectations. The studies are accompanied by neurological evidence which suggests that the cerebellum contains multiple forward and inverse models [4,5,6] and neuropsychological studies of patients with motor control deficits which suggest that the parietal lobes are involved in implementing forward models [7,8].

The concepts of forward models and inverse models have their roots in control theory where communication delays between a centralised control system, a motor system and sensory feedback render a sense-plan-act cyclic approach to action control inadequate. Application of the concepts to motor control is perhaps not surprising given that direct relationship between motor behaviour and action. However, the past 20 years has witnessed the increasing realisation that there is a second area of cognitive science where control is critical – namely in the generation and regulation of cognitive processes. Thus *cognitive control* is now recognised as a domain of research in its own right (see, e.g., [9]), paralleling the more classical domains of memory, perception, attention, reasoning, problem solving, etc. which populate the contents pages of standard cognitive textbooks. Two questions naturally arise. First, are there similarities between the problems of cognitive control and motor control? And second, if so, are the solutions apparently adopted in the case of motor control appropriate for the case of cognitive control? It is these questions which motivate the current work.

<sup>1</sup> Department of Psychological Sciences, Birkbeck, University of London, WC1E 7HX, UK. Email: [R.Cooper@bbk.ac.uk](mailto:R.Cooper@bbk.ac.uk).

## 2 THE PROBLEM OF COGNITIVE CONTROL

A well-established finding in the literature on speeded choice response tasks is that mean response time on error trials is generally shorter than on non-error trials and response time on trials immediately following an error is generally longer than on other non-error trials [10]. Following the initial report of this effect, subsequent studies have suggested that there is a gradient of response times, with RT decreasing within a run of correct trials until an error is made. RT then increases dramatically on the subsequent (post-error) trial, before gradually decreasing until another error is made [11]. Figure 1 illustrates the basic effect across five trials.

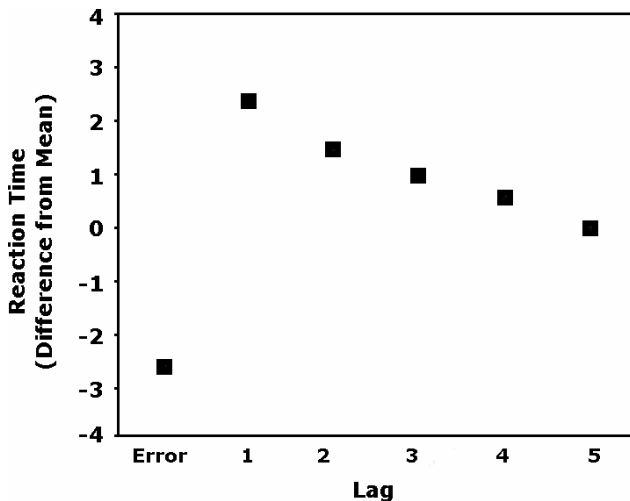


**Figure 1:** The effect of distance from error on RT over successive trials in a choice response task

What might lead to this effect? Botvinick and colleagues [12] suggest that it reflects the operation of a cognitive system that dynamically configures itself so as to optimise its performance. Essentially, in the absence of an error the cognitive system is held to adjust processing parameters so as to increase the speed with which a response is generated (e.g., by increasing the gain on one processing channel, or by decreasing the level of lateral inhibition operating across the system's output representations). If an error occurs, the system effectively recognises that it needs to take a more considered approach to its choice response and readjusts its processing parameters so as to slow processing. Botvinick et al consider three distinct reaction time tasks with similar characteristics where processing involves the resolution of conflict between competing responses (as in the standard choice reaction-time task), and argue for a general mechanism of *conflict monitoring* that applies across the tasks and has the control function described above.

Conflict monitoring and the subsequent adjustment of processing parameters in response to processing conflict reflects just one of a range of proposed cognitive control processes. Miyake and colleagues, for example, provide behavioural evidence based on a large individual-differences study for three specific cognitive control functions: task shifting, response inhibition and memory updating / maintenance [13]. These functions are not intended to be exhaustive. Thus other studies have suggested that dual-tasking might involve some additional,

separable, control process [14; see also 13]. Alternative partitionings of control functions are also available. Thus Stuss, Shallice and colleagues argue on the basis of the behaviour of neurological patients with lesions to different regions of prefrontal cortex for four identifiable control functions, which they label as task-setting, monitoring, energisation and attentiveness [15].

## 3 EXISTING COMPUTATIONAL ACCOUNTS OF COGNITIVE CONTROL

A critical difficulty relating to the studies of Miyake and colleagues and Stuss, Shallice and colleagues is that they lack operational accounts of the various control functions identified [16]. A strength of the Botvinick et al account [12], regardless of its veracity, is that it is instantiated in a set of computational simulations. These simulations stimulated a series of models of the potentially more general control function of *performance monitoring*, with the most recent being the ACC-RO model of Alexander and Brown [17]. We review each of these models as they represent two ends of a continuum where, at one end at least, concepts of prediction (and hence forward and inverse models) have been invoked.

The structure of one of the conflict monitoring models of Botvinick et al is shown in Figure 2. The model is of the well-known Stroop task, and is designed to capture the fact that the amount of Stroop interference (i.e. the difference in naming latency between incongruent and congruent Stroop trials) varies within subjects as a function of the proportion of trials within a block that are incongruent or congruent. When most trials are incongruent, subjects show less interference than when most trials are congruent [18]. This is captured within the model (which is based on an earlier, well-tested, model of Cohen and Huston [19]) through the conflict monitoring unit, which calculates a measure of conflict in the response layer and, dynamically adjusts gain on the task demand units. On incongruent trials, response layer conflict is high. This results in what the authors refer to as “a tightening of control”, via an increase in input to the colour-naming task demand unit (effectively increasing the gain of the non-dominant channel). This in turn leads to faster responses on incongruent trials.

The Botvinick et al account of conflict monitoring therefore does not make explicit use of forward or inverse models in the sense of the motor control literature. However, the models may be criticised because the consequence of response conflict differ across the three tasks which they consider. Thus, while in all cases conflict monitoring operates on response units (suggesting that it might be related to Miyake et al's response inhibition function, which those authors hold is reflected in Stroop task interference), in two of the cases the functional consequence of high response conflict is to increase attentional bias (i.e., the gain on one channel of attention), but in the third it functions by modulating the baseline activation of response units. A further problem is that, at least in the case of the Stroop model, it is unclear whether conflict can be plausibly controlled by amplifying task demand units. Other models of Stroop have used the same mechanism to account for task switching effects within the paradigm [20], yet these effects on most accounts are concerned with a cognitive control function that is distinct from that related to the resolution of conflict.

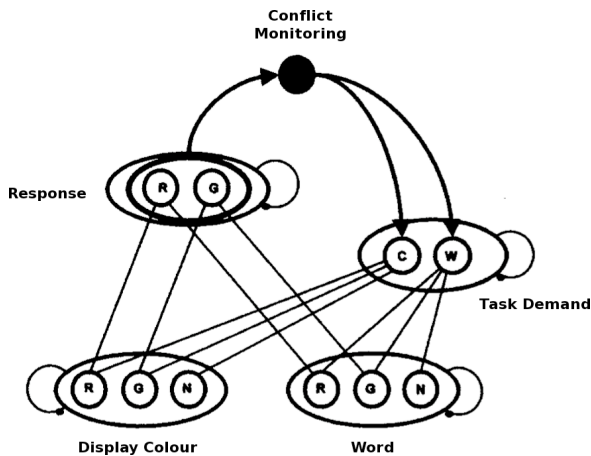


**Figure 2:** The Botvinick et al conflict monitoring model of the Stroop task.

A more recent model that builds on the ideas of Botvinick and colleagues and in some ways elucidates the above issues is the ACC-RO model of Alexander and Brown [17]. The model, which is illustrated in Figure 3, makes explicit use of a forward model to predict the outcome of a planned response. The response-outcome mapping (i.e., the forward model) is, it is argued, acquired through reinforcement learning. The prediction it provides is then compared with the action's actual outcome. Any discrepancy may be used as a control signal to adjust processing within the cognitive system.

Two features of the ACC-RO model warrant special attention. First, since the control signal is based not on conflict but on the mismatch between predicted and actual outcomes, the model provides a more general solution to the problem of cognitive control because it is not anchored to the type of interactive activation model considered by Botvinick et al. Second, it is conceivable that the control signal is used in different ways in different tasks. Within the model as it stands the control signal is a scalar value. This demands only a very impoverished forward model – possibly just a look-up table of actions and expected outcomes. The goal of cognitive control can then be phrased in terms of minimising this scalar value. Conceivably this may be achieved in different ways – increasing gain on one processing channel, reducing lateral inhibition between competing output nodes, allocating greater resources to a task, and so. In principle the preferred use of the control signal may be learned on a task-by-task basis. The downside of a scalar error signal however, is that it is necessarily non-specific – it can indicate sub-optimal configuration of the cognitive system, but it cannot differentiate between different causes for this sub-optimality.

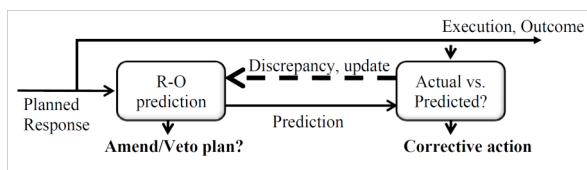


**Figure 3:** The performance monitoring ACC-RO model of Alexander and Brown.

## CONCLUSION

We have seen that some aspects of control theory, at least relating to forward models and the use of prediction, also have a potential place in cognitive control. Existing models that make use of this, however, adopt an extremely impoverished concept of a forward model. Whether this impoverished view is sufficient is the focus of current research.

## REFERENCES

- [1] D.M. Wolpert, Z. Ghahramani and M.I. Jordan. An internal model for sensorimotor integration. *Science*, **269**, 1880-1882 (1995).
- [2] D.M. Wolpert and Z. Ghahramani. Computational principles of movement neuroscience. *Nature Neuroscience*, **3**, 1212-1217 (2000).
- [3] D.A. Braun, A. Aertsen, D.M. Wolpert and C. Mehring. Learning optimal adaptation strategies in unpredictable motor tasks. *Journal of Neuroscience*, **29**, 6472-6478 (2009).
- [4] M. Kawato. Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, **9**, 718-727 (1999).
- [5] D.M. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, **11**, 1317-1329 (1998).
- [6] H. Imamizu, T. Kuroda, S. Miyauchi, T. Yoshioka and M. Kawato. Modular organisation of internal models of tools in the human cerebellum. *Proceedings of the National Academy of Science, USA*, **100**, 5461-5466. (2003)
- [7] A. Sirigu, J.R. Duhamel, L. Cohen, B. Pillon, B. Dubois and Y. Agid. The mental representation of hand movements after parietal cortex damage. *Science*, **273**, 1564-1568 (1996).
- [8] D.M. Wolpert, S.J. Goodbody and M. Husain. Maintaining internal representations: The role of the human superior parietal lobe. *Nature Neuroscience*, **1**, 529-533 (1998).
- [9] S. Monsell and J. Driver. (Eds.) *Attention and Performance XVIII: Control of Cognitive Processes*. MIT Press, Cambridge, MA (2000).
- [10] P. Rabbitt. Error and error-correction in choice response tasks. *Journal of Experimental Psychology*, **71**, 264-272 (1966).
- [11] D.R.J. Laming. *Information Theory of Choice-Reaction Times*. Academic Press, London, UK (1968).
- [12] M.M. Botvinick, T.S. Braver, D.M. Barch, C.S. Carter and J.D. Cohen. Conflict monitoring and cognitive control. *Psychological Review*, **108**, 624-652 (2001).
- [13] A. Miyake, N.P. Friedman, M.J. Wmerson, A.H. Witski and A. Howerter. The unity and diversity of executive functions and their contribution to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, **41**, 49-100 (2000).
- [14] R.H. Logie, G. Cocchini, S. Dela Sala and A.D. Baddeley. Is there a specific executive capacity for dual task coordination? Evidence from Alzheimer's Disease. *Neuropsychology*, **18**, 504-513 (2004).
- [15] T. Shallice, D.T. Stuss, T.W. Picton, M.P. Alexander and S. Gilligam. Mapping task switching in frontal cortex through neuropsychological group studies. *Frontiers in Neuroscience*, **2**, 79-85 (2008).
- [16] R.P. Cooper. Cognitive control: Componential or emergent? *Topics in Cognitive Science*. To appear.
- [17] W.H. Alexander and J.W. Brown. Computational models of performance monitoring and cognitive control. *Topics in Cognitive Science*. To appear.
- [18] J. Tzelgov, A. Henik and J. Berger. Controlling Stroop effects by manipulating expectations for color words. *Memory and Cognition*, **20**, 727-735 (1992).
- [19] J.D. Cohen and T.A. Huston. Progress in the use of interactive models for understanding attention and perception. In C. Umiltà and M. Moscovitch (eds.), *Attention and performance XV* (pp. 453-456). MIT Press, Cambridge, MA (1994).
- [20] S. Gilbert and T. Shallice. Task switching: A PDP model. *Cognitive Psychology*, **44**, 297-337 (2002).

# Where Do Rewards Come From?<sup>1</sup>

Richard L. Lewis<sup>2</sup> and Satinder Singh<sup>3</sup> and Andrew G. Barto<sup>4</sup>

**Abstract.** Reinforcement learning (RL) has achieved broad and successful application in cognitive science in part because of its general formulation of the adaptive control problem as the maximization of a scalar reward function. The computational RL framework is motivated by correspondences to animal reward processes, but it leaves the source and nature of the rewards unspecified. This paper advances a general computational framework for reward that places it in an evolutionary context, formulating a notion of an *optimal reward function* given a fitness function and some distribution of environments. Novel results from computational experiments show how traditional notions of extrinsically and intrinsically motivated behaviors may emerge from such optimal reward functions. In the experiments these rewards are discovered through automated search rather than crafted by hand. The precise form of the optimal reward functions need not bear a direct relationship to the fitness function, but may nonetheless confer significant advantages over rewards based only on fitness.

## 1 INTRODUCTION

In the computational RL framework [10], rewards—more specifically, reward functions—determine the problem the learning agent is trying to solve. Properties of the reward function influence how easy or hard the problem is, and how well an agent may do, but RL theory and algorithms are completely insensitive to the source of rewards (except requiring that their magnitude be bounded). This is a strength of the framework because of the generality it confers, capable of encompassing both homeostatic theories of motivation in which rewards are defined as drive reduction (as has been done in many motivational systems for artificial agents [7]), and non-homeostatic theories that can account, for example, for the behavioral effects of electrical brain stimulation and addictive drugs. But it is also a weakness because it defers key questions about the nature of reward functions.

Motivating the computational RL framework are the following correspondences to animal reward processes. Rewards in an RL system correspond to *primary rewards*, i.e., rewards that in animals have been hard-wired by the evolutionary process due to their relevance to reproductive success. In computational RL, they are thought of as the output of a “critic” that evaluates the RL agent’s behavior. Further, RL systems that form value functions, using, for example, Temporal Difference (TD) algorithms, effectively create *secondary* or *higher-order* reward processes whereby predictors of primary rewards act as

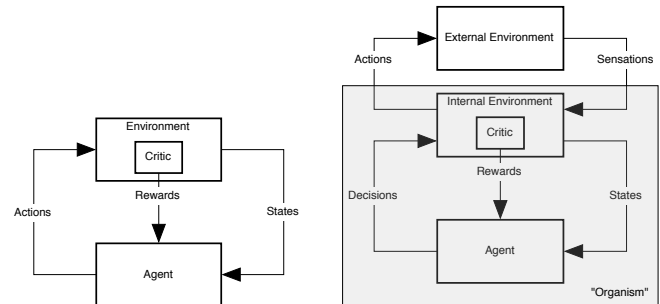


**Figure 1.** Agent-environment interactions in reinforcement learning; adapted from [1]. Left panel: Primary reward is supplied to the agent from its environment. Right panel: An elaboration in which the environment is factored into an internal and external environment, with all reward coming from the former.

rewards themselves. The learned value function provides immediate evaluations that are consistent with the infrequent evaluations of the hard-wired critic.

The result is that the local landscape of a value function gives direction to the system’s preferred behavior: decisions are made to cause transitions to higher-valued states. A close parallel can be drawn between the gradient of a value function and *incentive motivation* [3].

In the usual view of an RL agent interacting with an external environment (left panel of Figure 1), the primary reward comes from the external environment, being generated by a “critic” residing there. But as Sutton and Barto [10] and Barto, Singh, and Chentanez [1] point out, this is a seriously misleading view of RL if one wishes to relate this framework to animal reward systems.

In a less misleading view of this interaction (right panel of Figure 1), the RL agent’s environment is divided into external and internal environments. For an animal, the internal environment consists of the systems that are internal to the animal while still being parts of the learning system’s environment. This view makes it clear that reward signals are always generated within the animal, for example, by its dopamine system. Therefore, *all rewards are internal*, and the internal/external distinction is not a useful one (a point also emphasized by Oudeyer and Kaplan [5]). This is the viewpoint we adopt in this paper.

But what of a distinction between intrinsic and extrinsic reward? Psychologists distinguish between *extrinsic motivation*, which means doing something because of some specific rewarding outcome, and *intrinsic motivation*, which refers to “doing something because it is inherently interesting or enjoyable” [6]. According to this view, intrinsic motivation leads organisms to engage in exploration, play, and

<sup>1</sup> This paper was originally published as Singh, S., Lewis, R. L. and Barto, A. (2009) “Where Do Rewards Come From?”, *Proceedings of the Annual Conference of the Cognitive Science Society*, pp. 2601–2606.

<sup>2</sup> Department of Psychology, University of Michigan, Ann Arbor, email: rickl@umich.edu

<sup>3</sup> Division of Computer Science & Engineering, University of Michigan, Ann Arbor, email: baveja@umich.edu

<sup>4</sup> Department of Computer Science, University of Massachusetts, Amherst, email: barto@cs.umass.edu

other behavior driven by curiosity in the absence of explicit reinforcement or reward.

Barto et al. (2004) used the term *intrinsic reward* to refer to rewards that produce analogs of intrinsic motivation in RL agents, and *extrinsic reward* to refer to rewards that define a specific task as in standard RL applications. But the distinction between intrinsic and extrinsic reward is difficult to make precise. Oudeyer and Kaplan [5] provide a thoughtful typology. Space does not permit providing more detail, except to point out that a wide body of data shows that intrinsically motivated behavior does not occur because it had previously been paired with the satisfaction of a primary biological need in the animal's own experience [2]. That is, intrinsic reward is not the same as secondary reward. It is likely that the evolutionary process gave exploration, play, discovery, etc. positive hedonic valence because these behaviors contributed to reproductive success throughout evolution. Consequently, we regard intrinsic rewards in the RL framework as primary rewards, hard-wired from the start. Like any other primary rewards in RL, they come to be predicted by the value system. These predictions can support secondary reinforcement so that predictors of intrinsically rewarding events can acquire rewarding qualities just as predictors of extrinsically rewarding events can.

In short, once one takes the perspective that all rewards are internal (Figure 1), it is clear that the RL framework naturally encompasses and provides computational clarity to a wide range of reward types and processes, and thus has the potential to be a source of great power in explaining behavior across a range of domains. But fully realizing this scientific promise requires a computational framework for reward itself—a principled framework with generative power. Our main purpose here is to specify and illustrate a candidate for such a framework with the following desired properties:

### 1.1 Criteria for a Framework for Reward

1. The framework is *formally well-defined and computationally realizable*, providing clear answers to the questions of what makes a good reward and how one may be derived.
2. The framework makes *minimal changes to the existing RL framework*, thereby maintaining its generality.
3. The framework *abstracts away from specific mechanisms associated with RL agents*, such as whether their learning mechanisms are model-based or model-free, whether they use options or other kinds of richer internal structure, etc. But it is in principle powerful enough to exploit such agent structure when present.
4. The framework *does not commit to specific search processes* for finding good reward functions, but it does define and give structure to the search problem.
5. The framework *derives rewards that capture both intrinsic and extrinsic motivational processes*.

Taken together, these features of our framework distinguish it from other efforts aimed at deriving or specifying the form of reward functions (e.g., [8, 9, 4]). While these computational approaches are all valuable explorations of reward formulations, they still incorporate some notion of external reward, and are not concerned with explaining how their internal rewards come about. Closer to our aims is recent work by Uchibe and Doya [11] that proposes a specific mechanism—embodied evolution—for evolving reward. But this work is also still concerned with combining internal and external rewards, depending on specialized RL algorithms for guaranteeing that the asymptotic policy does not differ from the one implied by the external reward. The framework we propose here shares the

goal of providing an evolutionary basis, but dispenses with external rewards and seeks maximum generality in its formulation.

## 2 OPTIMAL REWARDS

Adopting an evolutionary perspective leads naturally to an approach in which adaptive agents, and therefore their internal reward functions, are evaluated according to their *expected fitness* given an explicit fitness function and some distribution of environments of interest. The fitness function maps trajectories of agent-environment interactions to scalar goodness values, and may take any form (including functions that are similar in form to discounted sums of external rewards).

### 2.1 Definition

More specifically we define the notion of optimal reward as follows. For a given RL agent  $A$ , there is a space of reward functions  $I_A$  that map agent internal state to a scalar primary reward that drives reinforcement learning. The nature of the internal state depends on the agent architecture. In general,  $A$  might include very specific commitments to particular learning algorithms and parameters so that the best  $I_A$  is partially determined by properties of the agent-as-learner. There is a distribution over Markov decision process (MDP; [10])<sup>5</sup> environments in some set  $\mathcal{E}$  in which we want our agents to perform well (in expectation). A specific reward function  $i_A \in I_A$  and a sampled environment  $E \in \mathcal{E}$  produces  $h$ , the history of agent  $A$  adapting to environment  $E$  using the reward function  $i_A$ . A given fitness function  $G$  produces a scalar evaluation,  $G(h)$  for all such histories  $h$ . The optimal reward function  $i_A^* \in I_A$  is the reward function that maximizes the expected fitness over the distribution over environments.

The formulation is very general because the constraints on  $A$ ,  $I_A$ ,  $G$ , and  $\mathcal{E}$  are minimal.  $A$  is constrained only to be an adaptive agent that uses a reward function  $i_A \in I_A$  to drive its search for control policies.  $G$  is constrained only to be a function that maps (finite or infinite) histories of agent-environment interactions to scalar goodness values. And  $\mathcal{E}$  is constrained only to be a set of MDPs, though the Markov assumption can be easily relaxed (we leave this to future work).

### 2.2 Regularities within and across environments

The above formulation essentially defines a search problem—the search for  $i_A^*$ . This search is for a primary reward function and is to be contrasted with the search problem faced by an agent during its lifetime, that of learning a good secondary reward function specific to its environment. Thus, our concrete hypothesis is (1) the  $i_A^*$  derived from search will capture physical regularities across environments in  $\mathcal{E}$  as well as complex interactions between  $\mathcal{E}$  and specific structural properties of the agent  $A$  (note that the agent  $A$  is part of its environment and is constant across all environments in  $\mathcal{E}$ ), and (2) the value functions learned by an agent during its lifetime will capture regularities present within its specific environment that are not shared across environments.

<sup>5</sup> An MDP is a mathematical specification of agent-environment interaction in which the environment can be in one of a fixed number of states, at each time step the agent executes an action from a fixed set of actions, and the action executed stochastically changes the state of the environment according to fixed transition probabilities.

### 3 TWO SETS OF COMPUTATIONAL EXPERIMENTS

We now describe a set of computational experiments in which we directly specify  $A$ ,  $G$ , and  $\mathcal{E}$ , and derive  $i_A^*$  via exhaustive search. These experiments are designed to serve three purposes. First, they will provide concrete and transparent illustrations of the basic framework above. Second, they will demonstrate the *emergence* of interesting reward function properties that are not direct reflections of the fitness function—including features that might be intuitively recognizable as candidates for plausible intrinsic and extrinsic drives in natural agents. Third, they will demonstrate the *emergence* of interesting reward functions that capture regularities across environments, and similarly demonstrate that value function learning by the agent captures regularities within single environments.

#### 3.1 Basic form of each experiment

Both experiments use a simulated physical space shown by the  $6 \times 6$  gridworld in Figure 3 (the arrows in that figure will be explained below). It consists of four subspaces (of size  $3 \times 3$ ). There are four movement actions, North, South, East and West that if successful move the agent probabilistically in the direction implied and if they fail keep the agent in place. The thick black lines in the figure represent barriers that the agent cannot cross, so that the agent has to navigate through gaps in the barriers to move to adjacent subspaces. The agent lives continually for its lifetime, i.e., the interaction is not episodic. Each experiment will introduce objects into the gridworld such as food or water or boxes. The internal state of the agent includes its location in the grid as well as other features relevant to each experiment. These features as well as other experiment-specific aspects (e.g., the fitness functions used) will be described in the appropriate sections below.

Our agent uses the  $\epsilon$ -greedy Q-learning [12] algorithm to learn how to act during its lifetime. This algorithm has three types of parameters: 1)  $Q_0$ , the initial Q-value function (we will use  $Q_0 = 0.0$  throughout) that maps state-action pairs to their long-term utility, 2)  $\alpha$ , the learning rate, and 3)  $\epsilon$ , the exploration parameter (at each time step the agent executes a random action with probability  $\epsilon$  and the greedy action with respect to the current Q-value function with probability  $(1 - \epsilon)$ ). At time step  $t$ , the current state is denoted  $s_t$ , the current Q-value function is denoted  $Q_t$ , the agent executes an action  $a_t$ , and the Q-learning update is as follows:

$$Q_{t+1}(s_t, a_t) = (1.0 - \alpha)Q_t(s_t, a_t) + \alpha[r_t + \gamma \max_b (Q_t(s_{t+1}, b))],$$

where  $r_t$  is the reward specified by the internal reward function  $i_A$  for the state  $s_{t+1}$ , and  $\gamma$  is a discount factor that makes immediate reward more valuable than later reward (we use  $\gamma = 0.99$  throughout). It is well known that the form of Q-learning used above will converge asymptotically to the optimal Q-value function and hence the optimal policy [12]. Thus, our agent uses its experience to continually adapt its action selection policy to improve the amount of discounted sum of reward it obtains in the future (remaining in its lifetime). Note that the reward function is distinct from the fitness function  $G$ .

The pseudo-code below describes how the *mean cumulative fitness* for a reward function  $i_A$  and algorithm parameters  $(\alpha, \epsilon)$  is estimated using simulations.

**for**  $i = 1$  **to**  $N$  **do**

    Sample an environment  $E_i$  from  $\mathcal{E}$ .

    In  $A$ , initialize Q-value function to zeros; set  $(\alpha, \epsilon)$

        Generate a history  $h_i$  over lifetime  $T$  for  $A$  and  $E_i$

        Compute fitness  $G(h_i)$

**end for**

$G(i_A) = \text{average of } \{G(h_1), \dots, G(h_N)\}$

The fitness of a reward function  $i_A$  is thus the maximum fitness returned over all  $(\alpha, \epsilon)$  pairs<sup>6</sup>.

#### 3.2 Hungry-Thirsty Domain: Emergent Extrinsic Reward for Water

In this experiment, each sampled environment has two randomly chosen special locations (held fixed throughout the lifetime of the agent); one where there is always food available and one where there is always water available. In addition to the movement actions, the agent has two special actions available: *eat*, which has no effect anywhere but when the agent is at the food location in which it causes the agent to consume food, and *drink*, which has no effect anywhere but at the water location in which it causes the agent to consume water. When the agent eats food it becomes not-hungry for one time step and then become hungry again. When the agent drinks water it becomes not-thirsty for a random period of time (when not-thirsty it becomes thirsty with probability 0.1 at each time step). Each time step the agent is not hungry, its fitness is incremented by one. There is *no* fitness directly associated with water at all. However being thirsty has a special effect; when the agent is thirsty its *eat* action fails. Thus the agent cannot just hang out at the food location and keep eating for at some point it will get thirsty and eating will fail. What is constant across environments is that there is food and there is water, that not-thirsty switches to thirsty with probability 0.1 and that being thirsty makes the agent incapable of eating. What varies across environments is the location of food and water.

Figure 2 presents some of our results. The fitness-based rewards assign a positive reward to the events that increment fitness (in this case not being hungry) and zero to everything else. The first interesting result (seen in the left panel) is that there are many reward functions that outperform the fitness-based reward functions throughout the lifetime (of 80,000 time steps) in terms of mean cumulative fitness. Careful inspection of the left panel also shows that some reward functions outperform the fitness-based reward functions early on in the lifetime but then are worse later on in the agent's life. We expect this to be a general phenomenon (and will study this in more detail in future work). The best reward function in our search space assigns a penalty of  $-0.2$  to the agent being hungry and thirsty, a smaller penalty of  $-0.1$  to the agent being hungry but not thirsty, a small positive reward of  $0.1$  to being not hungry but thirsty, and a large reward of  $1.0$  to not being hungry and not being thirsty. It is apparent that this reward function differentiates based on the thirst status of the agent. Thus, the search procedure has encoded into the reward function the emergent drive to drink water when thirsty. The Q-learning performed by the agent during its lifetimes learns a policy specific to the location of the food and water and takes the agent back and forth between food and water even though water does not directly contribute to cumulative fitness. This can be seen by inspecting Figure 3 that shows the policy learned by the agent (split into two panels; the right panel showing the actions when the agent is hungry

<sup>6</sup> Finding good reward functions for a given fitness function thus amounts to a huge search problem. In this paper, we conduct the search (mostly) exhaustively because our focus is on demonstrating the generality of our framework and on the nature of the reward functions found. However, there is structure to the space of reward functions (as we illustrate through our experiments) that we will exploit in future work to gain computational efficiency.

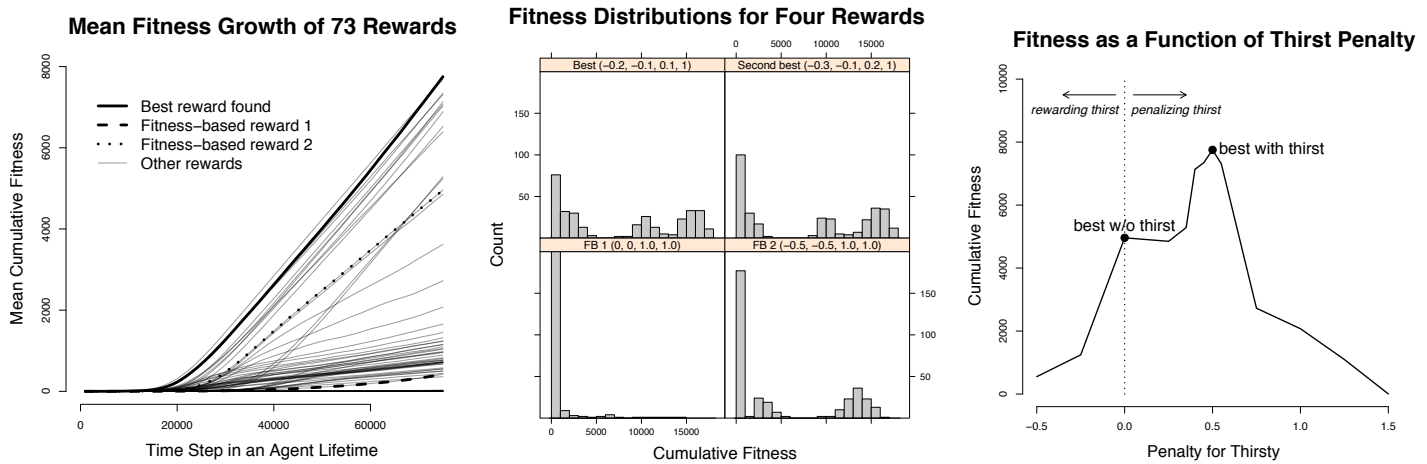


**Figure 2.** Results from Hungry-Thirsty Domain. See text for an explanation.

and thirsty and the left panel showing the policy when it is hungry and not thirsty)<sup>7</sup>.

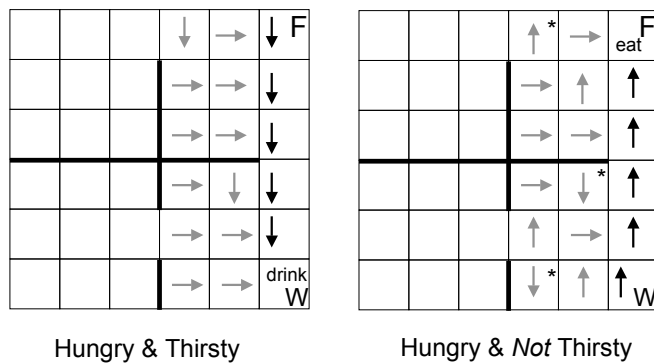


**Figure 3.** Policy for a single agent in Hungry-Thirsty Domain. See text for an explanation.

The middle panel in Figure 2 plots the distributions of cumulative fitness scores (at the end of the 80,000 time step lifetime) for 320 agents for each of four different reward functions (the two best rewards and the two fitness-based rewards). For all four rewards, the high variance and the multi-modal structure of the distributions is quite striking (and to us unexpected). The multimodal distribution of cumulative fitness also clearly shows that the fitness-based rewards do very poorly more than half the time.

Finally, the rightmost panel in Figure 2 shows how fitness is sensitive to the magnitude of the *penalty* that the reward functions provide

for being thirsty<sup>8</sup>. Each point along the horizontal axis corresponds to a subspace of reward functions that assign a specific penalty to being thirsty. The zero point on the horizontal therefore represents all those rewards that do not discriminate based on thirst, and the point labeled “best w/o thirst” indicates the fitness of the best-performing reward in that subspace. This reward corresponds exactly to the reward labeled “Fitness-based Reward 2” in the other panels. The point at the peak labeled “best with thirst” corresponds exactly to the reward labeled “Best reward found” in the other panels. This panel provides just one simple way of viewing the *fitness surface* associated with the reward space. It clearly indicates that not just any penalty for thirst will work well—only a relatively narrow and sharply peaked region of thirst penalty outperforms the best rewards that are insensitive to thirst.

### 3.3 Boxes Domain: Emergent Intrinsic Reward for Exploration and Manipulation

In this experiment, each sampled environment has two boxes placed in randomly chosen special locations (held fixed throughout the lifetime of the agent). In addition to the usual movement actions, the agent has two special actions: *open*, which opens a box if the agent is at the location of the box and has no effect otherwise, and *eat*, which has no effect unless the agent is at a box location, the box at that location is open, and there happens to be food in that box, in which case the agent consumes that food. An open box has 0.1 probability of closing at every time step. A closed box always has food. An open box never has food except for one time step after being opened from a closed state. Thus to consume food, the agent has to find a closed box, open it, and eat immediately in the next time step. When the agent consumes food it feels not hungry for one time step and its fitness is incremented by one. In addition to the location, the open/closed status of boxes is available to the agent as an observation at all times. The fitness based reward functions assign a positive reward for the fitness-inducing event, i.e., for being not-hungry and

<sup>7</sup> We only show the policy in the two subspaces containing the food and water because after learning the agent basically walks up and down the corridor connecting food with water and only falls off it due to random exploration. Thus the agent gets very little experience in the other subspaces and its policy there is mostly random. The policy off the path in the two right-subspaces is mostly correct (the exceptions are the three locations marked with a star).

<sup>8</sup> The thirst penalty for a given reward is simply the mean of the thirst penalty when hungry (the difference of the reward for hungry-but-not-thirsty and hungry-and-thirsty) and the thirst penalty when not hungry (the difference of the reward for not-hungry-and-not-thirsty and not-hungry-but-thirsty).

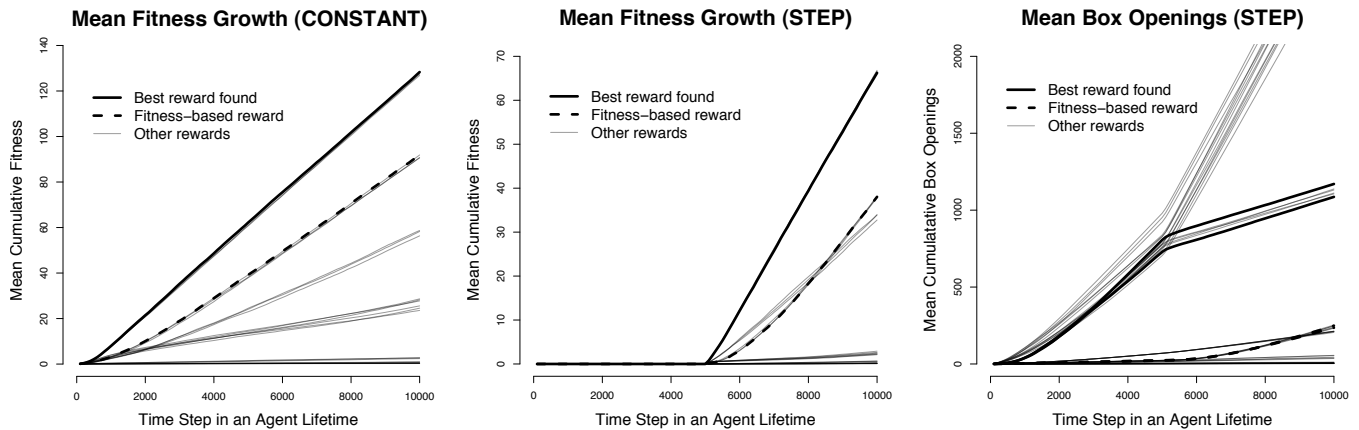


**Figure 4.** Results from Boxes Domain. See text for an explanation.

zero for everything else. What is unchanging across environments is the presence of two boxes and the dynamics of their open to close status. What changes across environments but held fixed within a single environment is the location of the boxes. Once again, this sets up the possibility for the search to hardwire across environment regularities into the reward function and then Q-learning to learn environment specific regularities.

In a different design from the first experiment, we ran this experiment under two different conditions. The first, called the *constant condition*, is one in which the food appears in closed boxes throughout the agent lifetime of 10,000 steps. The second, called the *step condition*, is one in which for the first half of the agent's lifetime there is never any food in any boxes and then after 5,000 time steps food always appears in a closed box. So in the step condition, there is no fitness to be obtained at all until after the 5,000<sup>th</sup> time step. The step condition simulates (in extreme form) a developmental process in which the agent is allowed to play in its environment for a period of time without having any fitness-inducing events present (in this case the fitness-events are good but in general there can also be bad ones that risk physical harm). Thus a reward function that is able to gain advantage through exposure to this first phase has to reward events that are intrinsically-motivating because they are by definition distal from any fitness-inducing events. Notice that this should happen in both the step and the constant conditions; it is just more stark and apparent in the step condition.

The top row of Figure 4 shows the average cumulative fitness as a function of time under the two conditions. As expected, in the step condition, there is no fitness under any reward function for the first 5,000 time steps. The best reward function for the step condition is as follows: not being hungry has a reward of 1.0, being hungry but with both boxes open has a slight reward of 0.1, one box open and hungry has a 0.0 reward, and both boxes closed and hungry has a slight penalty of -0.1. Clearly, the best reward function in our reward space has chosen to reward opening boxes. This makes the agent learn to open boxes during the first half of the step condition so that when the food appears in the second half the agent is immediately ready to exploit that situation. This is reflected in the constant slope from time 5,000 onwards of the average cumulative fitness curve of the best reward function. In contrast, the curve for the fitness-based reward function has a increasing slope because it has to learn to want

to open boxes from point 5,000 onwards. This difference is more clearly seen in the rightmost panel where we plot the average cumulative number of box openings for the different reward functions. The best reward function clearly opens the boxes far more often than the fitness-based reward function.

## 4 DISCUSSION AND CONCLUSIONS

We have outlined a general computational framework for reward that complements existing reinforcement learning theory by placing it in an evolutionary context—a context that clarifies and makes computationally precise the role of evolved reward functions in helping adaptive agents achieve fitness. We presented several computational experiments that serve to draw out and provide empirical support for two key properties of the framework.

First, multiple aspects of the domain may emerge to influence a single reward function. The combination of these multiple aspects is *implicit* in the form of the optimal reward function. Its precise properties are determined by the global goal of producing high fitness, but the relationship between the optimal reward function and fitness may be quite indirect. In the Hungry-Thirsty domain, two aspects of reward emerged, one related to food and hunger (directly related to fitness) and one related to thirst and water (not directly related to fitness). Both aspects were combined in a single function that represented a delicate balance between the two (Figure 2). In the Boxes domain, the optimal reward was related to food and hunger (directly related to fitness), and curiosity and manipulation of boxes (not directly related to fitness). The latter aspect of the optimal reward produced distinct *play and exploratory* behavior that would be thought of as intrinsically-motivated in the psychological sense. This was especially evident in the second condition of the Boxes experiment: during the first half of the agent's life time, no fitness-producing activities are possible, but intrinsically rewarding activities are pursued that have fitness payoff later.

The second key property of the framework is that two kinds of adaptation are at work: the local adaptation of the RL agent within a given environment, and the global adaptation of the reward function to both a population of environments and the structure of the agent itself. The two kinds of adaptation are apparent in both experiments. In the Hungry-Thirsty domain, each agent benefits from the

regularity across environments that drinking water ultimately helps it to achieve fitness—a regularity captured in the optimal (primary) reward function. Each agent also learns the specific locations of water and food sources in a given environment and good navigation patterns between them—regularities captured in the value functions learned by RL (Figure 3). Similarly, in the Boxes domain, each agent benefits from the regularity across environments that food is to be found in boxes, but also learns how to navigate to the specific locations of boxes in a given environment.

These are initial results in simple domains but they suggest several promising parallel avenues for future research. It is especially important to explore environments in which fitness selects for rewards that yield a diverse set of tasks and behaviors—environments in which the traditional notion of an external reward function provides the least guidance. It is also clear that the search for good reward functions represents a major computational challenge. Further work is required to develop more efficient approximate search algorithms that exploit the structure of this space, while still preserving its potential richness.

## ACKNOWLEDGEMENTS

Satinder Singh was supported by AFOSR grant FA9550-08-1-0418 and by NSF grant IIS 0905146. Richard Lewis was supported by ONR grant N000140310087 and by NSF grant IIS 0905146. Andrew Barto was supported by AFOSR grant FA9550-08-1-0418. Any opinions, findings, conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the sponsors.

## REFERENCES

- [1] A. G. Barto, S. Singh, and N. Chentanez, 'Intrinsically motivated learning of hierarchical collections of skills', in *Proceedings of the International Conference on Developmental Learning (ICDL)*, (2004).
- [2] E. L. Deci and R. M. Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior*, Plenum Press, N.Y., 1985.
- [3] S. M. McClure, N. D. Daw, and P. R. Montague, 'A computational substrate for incentive salience', *Trends in Neurosciences*, **26**, 423–428, (2003).
- [4] A. Ng, D. Harada, and S. Russell, 'Policy invariance under reward transformations: Theory and application to reward shaping', in *Proceedings of the Sixteenth International Conference on Machine Learning*, Morgan Kaufmann, (1999).
- [5] P.-Y. Oudeyer and F. Kaplan, 'What is intrinsic motivation? A typology of computational approaches', *Frontiers in Neurorobotics*, (2007).
- [6] R. M. Ryan and E. L. Deci, 'Intrinsic and extrinsic motivations: Classic definitions and new directions', *Contemporary Educational Psychology*, **25**, 54–67, (2000).
- [7] T. Savage, 'Artificial motives: A review of motivation in artificial creatures', *Connection Science*, **12**, 211–277, (2000).
- [8] J. Schmidhuber, 'A possibility for implementing curiosity and boredom in model-building neural controllers', in *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pp. 222–227, Cambridge, MA, (1991). MIT Press.
- [9] S. Singh, A. G. Barto, and N. Chentanez, 'Intrinsically motivated reinforcement learning', in *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, Cambridge MA, (2005). MIT Press.
- [10] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.
- [11] E. Uchibe and K. Doya, 'Finding intrinsic rewards by embodied evolution and constrained reinforcement learning', *Neural Networks*, **21**(10), 1447–1455, (2008).
- [12] C. J. C. H. Watkins, *Learning from Delayed Rewards*, Ph.D. dissertation, Cambridge University, Cambridge, England, 1989.
- [13] A. P. Wolfe, T. Tran, and A. G. Barto, 'Evolving reward and initial value functions', Technical report, University of Massachusetts, Department of Computer Science, Amherst, MA, (2008).

# Social Cognition: Views from developmental and comparative psychology, and robotics on the role of emotion in joint attention.

**Prof Kim A. Bard**  
Portsmouth University

## Abstract

Most human infants, by 12 months of age, have developed a suite of triadic engagement skills, specifically, joint attention, social referencing, and intentional communication (e.g., Russell et al., 1997). Developmental psychologists emphasised the cognitive underpinnings of how infants become able to coordinate their engagement with people and their engagement with objects to form the referential triangle of infant- mother-object (triadic) engagement (e.g., Bard, 1992). Developmental psychology provided the conceptual frame for many studies of comparative cognition, and some comparative psychologists have asserted that this suite of triadic engagement skills specifies the beginning of the uniquely human 'social cognitive revolution'. New evidence emerging from the FEELIX GROWING project (Canamero, 2008; Hiolle et al., 2009) highlights the importance of emotion (e.g., Murray et al., 2009), and suggests a simple learning system can account for the development of triadic engagements (e.g., Boucenna et al, 2010). Evidence emerging from my developmental comparative studies (e.g., Bard & Leavens, 2009) and those of my colleagues (e.g., Leavens et al., 1996, 2005) indicate that emotion plays a crucially important role in the development of social cognition in young apes, as well (Bard, in press; Bard et al., 2005; Leavens et al., in press).

## References

Bard, K.A. (1992). Intentional behavior and intentional communication in young free-ranging orangutans. *Child Development*, 63, 1186-1197.

Bard, K.A. (in press). Emotional engagement: how chimpanzee minds develop. To appear in de Waal, F. & Ferrari, P. (eds), *The Primate Mind*. Cambridge, MA: Harvard University Press.

Bard, K.A. & Leavens, D.A. (2009). Socio-emotional factors in the development of joint attention in human and ape infants. In L. Roska-Hardy & E. Neumann-Held (Eds.), *Learning from animals? Examining the nature of human uniqueness* (pp. 89-104). London: Psychology Press.

Bard, K.A. Leavens, D.A., Custance, D., Van't Hof, M., Keller, H., Benga, O., & Sousa, C. (2005). Emotion cognition: Comparative perspectives on the social cognition of emotion. *Cognition, Creier, Comportament (Cognition, Brain, Behavior)*, Special Issue: "Typical and atypical development", VIII, 351-362.

Boucenna, S., Gaussier, P., Hafemeister, L., & Bard, K. (submitted, Feb 2010). Social referencing. To appear in *From Animals to Animats 11: Eleventh international conference on simulation adaptive behavior*, Lecture notes on Computer Science, Springer.

Canamero, L. (2008). Development within the "Complete Creature" Paradigm. *IEEE AMD Newsletter - The Newsletter of the Autonomous Mental Development Technical Committee*, Vol. 5, No. 2, November 2008.

Hiolle, A., Bard, K.A., & Canamero, L. (2009). Assessing human reactions to different robot attachment profiles. *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication Toyama, Japan, (RO-MAN 09)*, Sept. 27-Oct. 2, 2009, p.251- 256.

Leavens, D.A., Bard, K.A. & Hopkins, W.D. (in press). BIZARRE chimpanzees do not represent "The Chimpanzee". Commentary on target article by Heinrich, Heine & Norenzayan, 'The weirdest people in the world?', *Behavioral and Brain Sciences*, 33(1).

Leavens, D.A., Hopkins, W.D., & Bard, K.A. (1996). Indexical and referential pointing in chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, 110, 346-353.

Leavens, D.A., Hopkins, W.D., & Bard, K.A. (2005). Understanding the point of chimpanzee pointing: Epigenesis and ecological validity. *Current Directions in Psychological Science*, 14, 185-189.

Murray, J. C., Canamero, L., Bard, K. A., Davila Ross, M., and Thorsteinsson, K. (2009). The Influence of Social Interaction on the Perception of Emotional Expression: A case study with a robot head. In J.H. Kim et al (eds); *FIRA 2009, Lecture Notes in Computer Science 5744*, pp 63-72. Heidelberg, Berlin:Springer-Verlag.

Russell, C.L., Bard, K.A., & Adamson, L.B. (1997). Social referencing by young chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, 111, 185-193.

# **Biological and cultural foundations of human language: Insights from computer simulations.**

**Prof Nick Chater**  
**University College London**

## **Abstract**

It has been standard in linguistics to argue that language acquisition is so difficult that it is only possible if most linguistically interesting structure is embedded in a genetically-specified “language module,” “language instinct,” or “universal grammar.” But how could a biological basis for language have evolved, in the absence of a prior linguistic environment which could shape biological evolution. A natural, and popular, assumption is that the biological evolution of a language faculty must have co-evolved with the cultural evolution of language itself. Here I report work with Morten Christiansen, Florencia Reali, Andrea Baronchelli and Romualdo Pastor-Satorras, which places strong limitations on such co-evolution. Indeed, these simulations suggest that the traditional conception of a “universal grammar” can be ruled out on evolutionary grounds; and may have broader indications for claims that humans have biologically based adaptations to other aspects of culture.

# The role of vision and attention in language processing

**Prof Antje S. Meyer**

**School of Psychology, University of Birmingham, and  
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands**

## **Abstract**

In recent years, psycholinguistic theories have become increasingly concerned with the relationships between domain-specific linguistic and general cognitive processes. In my talk I will (a) discuss some of the relationships between vision, attention, and linguistic processing, and (b) demonstrate how eye tracking can be used to study these relationships.

In the first part of the talk I will review studies of spoken language comprehension demonstrating that the listener's eyes tend to be drawn to objects mentioned or implied in spoken utterances. I will propose that listeners aim to generate conceptual representations that, as much as possible, integrate visual and auditory information, that this process is facilitated by directing visual attention at the named or implied objects, and that the direction of visual attention is reflected in the eye movements.

In the second part of the talk I will review studies of language production demonstrating that speakers describing displays or events tend to fixate upon each object they name, and that their gaze remains focussed on the object until they have retrieved the sound form of its name. I argue that the tight link between eye gaze and speech output and the long gazes to the objects arise because directing one's visual attention at an object facilitates not only the recognition of the object but also the retrieval of associated information, including the object name. The easiest way of directing one's attention to an object is to fixate upon it.

In final part of the talk I will summarise some of the strengths and weaknesses of eye tracking as a tool for psycholinguistic research and suggest directions for future research.

# **Causal Networks in Neural Systems: from Brain-Based Devices to Consciousness**

**Dr. Anil Seth,**

School of Informatics, University of Sussex

## **Abstract**

Neurons engage in causal interactions with one another and with the surrounding body and environment. Neural systems can therefore be analyzed in terms of causal networks, without assumptions about information processing, neural coding, and the like. Causal networks therefore provide a lingua franca for translating insights obtained from computer simulations or robots into predictions and interpretations relevant to biological systems. My talk will be in three parts. First, I will describe the theoretical basis for ‘causal network analysis’ which rests on a combination of time-series analysis (“Granger causality”) and network theory. I will then describe its application to a series of “brain-based devices”, drawing implications for possible causal pathways in the hippocampus and for the relation between synaptic plasticity and behavioural learning. Finally, I will explore how causal networks shed new light on the possible biological mechanisms underlying conscious experience, with reference to the nascent project of ‘artificial’ or ‘machine’ consciousness.

## **References**

- Seth, A.K. (2009). The strength of weak artificial consciousness. *International Journal of Machine Consciousness*.
- Seth, A.K. (2008). Causal networks in simulated neural systems. *Cognitive Neurodynamics*, 2:49-64.

# Why Can't We Build Robots that are as Clever as Crows?

**Prof Murray Shanahan**  
Imperial College London,

## **Abstract**

In recent years the cognitive prowess of corvids has been amply demonstrated. Despite their evolutionary distance from primates and the correspondingly different organisation of their brains, corvids are capable of tool manufacture and use, physical and causal cognition, social cognition and deception, and mental time travel. But how are these capacities realised in their brains? A satisfactory answer to this question would surely meet the criterion of implementability. If we really knew what makes a crow tick, we would be able to build a clockwork crow. Conversely, if we could build a clockwork crow, then we might be in a position to advance hypotheses about what makes a real crow tick. But today's robots fall a long way short of corvid-level intelligence. In this respect, the prospects for AI-inspired biology look as forlorn as those for biologically-inspired AI. Perhaps the way to fill the explanatory hole is to look to the emerging field of neurodynamics for the right theoretical vocabulary.