

Self-directed agents

W.D. Christensen and C.A. Hooker¹

Abstract

In this paper we outline a theory of the nature of self-directed agents. On our account what is distinctive about self-directed agents is that they are able to anticipate interaction processes and evaluate their performance. This allows self-directed agents to modify their behaviour context sensitively so as to improve the achievement of goals, and in certain instances construct new goals. We contrast self-directedness with reactive action processes that are not modifiable by the agent, though they may be modified by supra-agent processes such as populational adaptation or external design. Self-directedness lies at the nexus of a set of issues concerning the evolution and nature of intentionality, intelligence and agency. It provides the core of a biologically grounded account of intentional agency.

1 Introduction

In this paper we outline a theory of the nature of self-directed agents. What is distinctive about self-directed agents is their ability to anticipate interaction processes and to evaluate their performance, and thus their sensitivity to context. They can improve performance relative to goals, and can, in certain instances, construct new goals. We contrast self-directedness with reactive action processes that are not modifiable by the agent, though they may be modified by supra-agent processes such as populational adaptation or external design.

Self-directedness lies at the nexus of issues concerning the evolution and nature of intentionality, intelligence, and agency. It provides some insight into the evolution of intelligence because it helps explain how organisms are able to manage variable interaction processes, e.g. a hunting strategy that varies with prey type, ground condition, and hunger level. Simple self-directed organisms like bumblebees manage variability in one or a few dimensions. They are able to track changes in the types of flowers that are yielding nectar by evaluating the outcome of flower visits using a gustatory reward signal, and learn to anticipate which flower types have reliable nectar yields. In more complex forms of self-directedness the variability may be in many dimensions, and effective management can require a form of learning we term *open problem solving*. Open problems occur where the agent initially lacks the ability to identify or act upon the key factors for producing a solution, and must discover the relevant factors and the actions that influence them through extended interactive learning. Skill formation is a form of open problem solving, and skilled activities such as hunting count as relatively sophisticated forms of self-directedness.

¹ Addresses: Wayne Christensen, Philosophy, School of Liberal Arts, University of Newcastle, Callaghan, Australia 2308. Email: plwdc@alinga.newcastle.edu.au. Cliff Hooker, School of Liberal Arts, University of Newcastle, Callaghan 2308, NSW, Australia. Email: plcah@alinga.newcastle.edu.au. We would like to thank Mark Bickhard, John Collier and Bill Herfel for constructive discussions. Jill McIntosh made numerous editorial suggestions which have greatly improved the clarity of the paper and some of the content. CAH thanks the Philosophy Departments at Durham University, UK and University of Western Ontario, Canada, for generous hospitality during part of the preparation of this paper.

At the high end of the spectrum are highly sophisticated and open-ended human cognitive abilities such as commanding a warship, starting a business, or conducting scientific research.

This account has widespread implications for understanding the nature of intentionality. First, it carries a general commitment to pursuing a dynamically situated agent-oriented approach, grounded in biologically realistic problems. Self-directedness is an ‘information hungry’ form of adaptiveness, in roughly the sense of Clark (1997, ch.10). Self-directed agents acquire information from the environment as part of the process of forming anticipations.² Understanding this process bears on understanding information content, because information must come in a form that the agent can use to modify action, and integrate with other information. We will discuss these issues and suggest that the currently popular teleosemantic theory of content does not satisfy the requirements on the nature of information posed by the kinds of processes with which we are concerned. Teleosemantic content is defined in terms of selection history, and therefore cannot be used by the agent because typical biological agents do not have information about their selection history. As an alternative we will propose an interactivist-constructivist account of intentionality that relates information content to action.³ This allows us to understand how agents can use and process information.

Second, this approach introduces a rich conception of the higher order cognitive processes involved in intelligence and agency. Affect processes, like hunger and thirst inducement and satiation, supply the norms required for evaluating interaction, and are thereby key factors in shaping the goals and learning processes of the agent. Agents may learn relationships among their basic affect conditions and also construct new, derived activity norms from them, leading to the development of a complex normative array for evaluating and guiding action (see §2). Self-directedness also involves the integration of information from multiple sources to focus action into coordinated activities. This can include resolving action conflict when there are several possible actions that are mutually exclusive or have antagonistic effects, learning about the relations between actions and outcomes, and planning ahead to achieve particular outcomes. Understanding these processes can also provide some insight into how diverse sources of information can be combined to form an overall situational awareness.

Third, it is plausible that the learning processes that are involved in strong forms of self-

² Forming anticipations sounds cognitively sophisticated but, as in the case of bee foraging we discuss below, it can occur through quite simple processes like operant conditioning. The ability to form anticipations is most likely to have arisen in animal evolution through the specialisation of more general capacities of neural systems for experience-based modification. Even the simplest neural systems, such as the two-neuron tentacle withdrawal reflex in coelenterates, are capable of habituation, so basic anticipative abilities like event expectancy and action priming are not too difficult to achieve. For a discussion of neural mechanisms for anticipative learning see Montague and Sejnowski 1994.

³ Christensen and Hooker 2000 develops an account of interactivist-constructivism (I-C) as an approach to understanding intelligence. See also Bickhard and Terveen 1995; Christensen 2000, ch.1. I-C is Piagetian in spirit, though not in detail. It emphasises the embodiment of intelligence, and has philosophical connections with pragmatism and aspects of phenomenology. Perhaps more importantly, though, I-C is designed to articulate a perspective on cognitive science and philosophy of mind that reflects contemporary research focussed on dynamical interaction and development. See e.g. Brooks 1991, Edelman 1987, Glenberg 1997, Hutchins 1995, Karmiloff-Smith 1992, Lakoff 1987, Pfeiffer and Sheier 1999, Quartz and Sejnowski 1997, Smith and Thelen 1993.

directedness play a central role in the formation of cognitive representations and concepts. Self-directed agents need to learn what affects success and failure. Part of this involves differentiating specific states-of-affairs, objects and object types. With experience, the concepts of an agent gain increasing definition and richness as the agent discovers more of the interaction characteristics of these entities. Moreover, the agent's ability to use concepts becomes increasingly flexible and open-ended as the agent gains greater interaction skills and is thus able to appropriately utilise concepts in a range of contexts.

All this adds up to a graded multi-dimensional conception of intentional agency that contrasts with the currently common one-dimensional conception of intentionality as a capacity for reference modelled on human language use. The more complex conception based on self-directedness provides a richer framework for understanding the evolution and development of intentionality and intelligence.

2 The envelope and the matrix: some general dynamical issues for understanding adaptive agents

2.1 How not to study the evolution of mind: matching cognitive and evolutionary modules.

One of the key challenges for evolution of mind research is developing an adequate approach to grappling with the overlap of the biological and cognitive domains. The problem is that biology, cognitive science, and philosophy employ very different methods, theories, and concepts. In certain respects the most obvious strategy for solving this problem is to find a direct association between entities postulated by theories in one domain and entities postulated by theories in another. And the simplest way to do this is by demonstrating that the functional modules in one domain turn out to be functional modules in another. Thus, evolutionary psychology compartmentalises the mind into a suite of 'domain specific computational modules', such as for mate selection and detection of social cheating, that are assumed to be heritable traits, and then speculates about the circumstances under which these putative traits might have been favoured by evolution. Similarly, though at a more general level, teleosemantics attempts to characterise representation as a kind of evolutionary function. However, the theoretical assumptions upon which such a unification strategy rests are controversial. Both evolutionary psychology and teleosemantics hinge on linking adaptationist neo-Darwinism with representationalist cognitive science.⁴ These theories make the job seem easy because they employ strong modularity assumptions, encouraging the idea that there may be a straightforward cross-theory mapping. Unfortunately adaptationism and representationalism have each come under strong criticism, not least because of their functional modularity assumptions.

In particular, it has been argued that the modularities assumed by adaptationism and representationalism characteristically tend to neglect interaction and development. In biology, the localisation of heredity to genes has been challenged by developmental approaches, which stress that heredity depends on the organisationally distributed dynamical processes of

⁴ See Stotz and Griffiths (in press) for a critical analysis of evolutionary psychology that makes this case.

development.⁵ Likewise, the assumption in cognitive science that adaptive behaviour is mediated by representations, forming a categorically distinct set of entities uniquely associated with intelligence, has been challenged by developmental, dynamical systems and behaviour-based robotics approaches which stress that adaptive behaviour is generated by organisationally distributed interaction processes.⁶ The similarity of these critiques is striking, especially given that the disciplinary contexts are quite different.

If, as evidence suggests, distributed dynamical processes play a fundamental and widespread role in adaptiveness and cognition, and orthodox approaches neglect this, two things follow. First, theories and concepts for understanding both biological and cognitive phenomena must explicitly recognise holistically structured dynamical relations. Second, where functional localisation is attributed, it must be given detailed context-specific justification.

It is here that the problems with drawing a general link between adaptationism and representationalism really bite. Neither adaptationism nor representationalism respect these two conditions. They can make little sense of holistically structured dynamical relations because the entire thrust of both theories is to localise functionality to particular structures: adaptations in the first case and representations in the second. Moreover, they both assume functional modularity as an a priori given, or at most as justified on very general grounds. But such generality is only achieved by neglecting the biological mechanisms involved, and unfortunately the nature of those mechanisms can make a large difference to the kinds of functional organisation that actually occur.⁷

Many people leap to the assumption that, since mind has evolved, it must be composed of adapted functional units. Hence the search for a basic cognitive ‘toolkit’, as some evolutionary psychologists refer to it. However the inference from evolution to functional units is faulty. Cognition doesn’t have to be composed of discrete functional units in order to evolve; selection may instead adjust the parameters of an integrated neuro-hormonal developmental process (cf.

⁵ E.g. Griffiths and Gray 1994, Jablonka and Lamb 1995, Oyama 1985.

⁶ E.g. Beer 1995, to appear, Brooks 1991, Hendriks-Jansen 1996, Smith and Thelen 1993, van Gelder 1995, 1998.

⁷ For example, proponents of evolutionary psychology argue that cognition will be highly modularised because specialised cognitive modules will tend to out-compete generalist cognitive modules. In a similarly general way Millikan argues that traits whose function is to represent will be produced by evolution because other traits need to correlate their activity with the environment in order to function properly. Both of these arguments are seriously weakened by the fact that they make no reference to the underlying biological mechanisms involved. Modularity has the strengths and weaknesses characteristic of any specialisation. Without information about system possibilities and costs in relation to required tasks it is impossible to specify what the possibilities and trade-offs are with respect to various modularising schemes. (Cherniak 1986 provides an illuminating general discussion of these tradeoffs for human memory: too little compartmentalisation and it takes too long to search a compartment, too much compartmentalisation and it takes too long searching for the relevant compartment.) The problem for evolutionary psychology is that without more information about neural architectural possibilities, niche characteristics and developmental processes, it is simply impossible to specify what the possibilities and trade-offs are with respect to cognitive modularisation. The problem for Millikan is that without a more detailed account of how the interaction processes are organised it is impossible to specify in any detail how system-environment correlations are generated and whether they are mediated by structures appropriately thought of as representations. We will discuss this further in the next section.

Kauffman 1993). Exactly how much modularity actually occurs in cognition is an empirical question, but we suggest that the evidence for distributed processes in interaction and development is strong enough to justify a very different project for understanding the evolution of mind. As Karmiloff-Smith (1992) argues, the issue for this kind of project is understanding how evolutionary biases and environmental interaction act in concert to shape the development of cognition.⁸ From this perspective, understanding the evolution of mind is not a matter of finding a mapping between evolutionary and cognitive units; it is a matter of understanding the various factors and complex interactions that shape the developmental processes of cognitive agents.

2.2 Integrated adaptive agency

Our strategy is not to find cognitive units that are also adaptations-- it is to piece together some of the factors that are likely to have played a role in shaping the phylogeny of intelligent organisms. Our account of self-directed agents is designed to serve as a synthetic model for relating the diverse conceptual and empirical issues that bear on intelligence. Moreover, the place to start is not with functional units but with whole systems and processes. Our account of self-directedness begins with the concepts of **performance envelopes** and **norm matrices**. These concepts provide a basis for understanding organisms as dynamically integrated systems, and help illuminate some of the adaptive issues that underlie the evolution of intelligent agents.

The concept of performance envelopes is useful for understanding the dynamical way that organisms survive and reproduce as complete integrated systems. Organisms clearly have functionally specialised components, however it doesn't follow that the overall adaptiveness of the organism should be thought of as a sum of discretely adaptive individual traits. In fact, overall adaptiveness is a complex nonlinear product of the interactions of many factors. These factors include at least the following general kinds:

- (i) Gross performance parameters of ecological interaction (e.g. running speed, sensory acuity).
- (ii) Fundamental systemic processes of the organism (e.g. cellular metabolism, development).
- (iii) Specialised component structures and sub-systems (e.g. heart, lungs, muscles).
- (iv) Factors that differentially affect development (e.g. DNA-protein synthesis, developmental molecular-cellular interactions, environmental resources).

The relations amongst these kinds of factors are many-many: a given gross performance parameter such as running speed will be influenced by many systemic processes, many structures and subsystems, and many developmental factors; a given systemic process, such as cellular metabolism, will play a role in many performance parameters, involve many components, and be affected by many developmental factors; and so on.

Thus, it is important to understand how adaptive systems perform as integrated systems under a potentially open-ended range of conditions. Performance envelopes provide a way of describing this: a performance envelope is determined by the interrelationships of a number of key system parameters. For example, we can characterise the performance envelope of human hand-eye coordination with respect to object manipulation. Given the broad parameters of upper body

⁸ For similar arguments see Griffiths 1997, Stotz and Griffiths in press.

strength, humans can lift objects within a certain weight range. This range affects what a human can do with an object. A human can hurl a small object a considerable distance, but as weight goes up, the distance the object can be thrown decreases, until a point is reached where the object can barely be lifted. With stereoscopic vision and fine motor control, humans can perform finely structured actions with objects, but the degree of control decreases as object size, temperature, slipperiness, or weight, increases. Thus, the interrelations amongst the parameters of human hand-eye coordination establish a performance envelope for object manipulation that determines the kinds of object manipulation tasks that humans can perform.

Note, however, that a performance envelope is *not* the same kind of thing as a task specification. The human performance envelope for object manipulation encompasses an open-ended range of tasks, ranging from peeling fruit to throwing rocks at wild dogs, making stone tools, playing musical instruments, painting, writing, and driving cars. In contrast, a task specification measures the performance of a system against a specific kind of task, such as stone tool making. This difference is important because in certain situations performance envelopes are more adaptively important than task specifications, particularly for understanding adaptive change. In this respect, it is worth observing that the etiological theory of functions is a theory of task specification: proper function is performance of the task that led to selection for the trait.

2.3 Performance norms for integrated agents

The performance envelope that is of most fundamental significance for adaptive systems is the one that corresponds to the integrity of the system itself. This is the system's **viability envelope**. We call this condition **autonomy** and have analysed it at some length elsewhere.⁹ Here we focus on using it to understand adaptive agents as integrated systems. Autonomous systems are self-regenerating (or 'self-governed') in the sense that they interactively contribute to the conditions required for their own existence. They engage in interaction processes that acquire resources from the environment, and transform these resources into the energy and infrastructure of the system itself, regenerating the whole, including the interactive capacities themselves. In this respect, autonomous systems are distinct from entities such as rocks which are merely passively internally stable (if a rock is damaged it won't perform work to reform itself), and from gases which are wholly externally stabilised by their environment. Autonomous systems include living cells, multicellular organisms, species and cities.

The key feature of autonomous systems is that they are composed of networks of interdependent processes whose integrated activity is self-generating. Thus, the **viability envelope** of the system is the range of conditions under which the process network constitutive of the system is self-generating. This process interdependence provides a way of understanding adaptive norms, because for the whole system to be self-generating its process activity must meet coordination requirements, and these coordination requirements act as the constraints, or norms, that determine success and failure for the system. For instance, a fundamental process involved in the viability of a cheetah is cellular metabolism, which imposes many requirements that must be

⁹ See Christensen and Hooker 2000, forthcoming, Christensen and Bickhard in press. A reasonably comprehensive discussion of autonomy is Christensen, Collier and Hooker, 'Autonomy', which forms ch.2 of Christensen's PhD thesis and is available at: <http://www.newcastle.edu.au/departments/pl/staff/WayneChristensen/dissertation.htm>

satisfied by other processes in the cheetah, such as a supply of oxygen generated by breathing and a supply of nutrients generated by hunting and feeding. Thus, a primary normative standard that determines success or failure for a cheetah's autonomic activity is adequate oxygen supply, and a primary normative standard for its hunting activity is the adequate supply of nutrients coordinated with the locations, quantities and timing of metabolic requirements.

This account of norms has some significant advantages over the standard etiological account. In particular, because it takes into account the overall organisation of the system and isn't tied to task specifications derived from previous behaviour, it allows us to understand adaptive change. In other words, we can understand how a change in an adaptive system's activities that has no precedent might still be considered normatively good or bad. This is important with respect to understanding evolutionary processes in general, but it is especially crucial for understanding intelligent agents because it allows us to characterise the norms that apply to choice. In particular, the norms that matter to an agent faced with a choice situation are those that concern its possibilities for action in the present circumstances and their likely outcomes, not what it or its ancestors did in the past.

We can illustrate this intuitively in terms of managing a business. Gerry's Cleaner Crisper Laundromat business has expanded slightly since Gerry hired Sandra to assist with dry cleaning and pressing clothes whilst he deals with customers, monitors the washing machines, performs minor tailoring and does the accounting. However, after Gerry hires Sandra, the government introduces a new tax system that requires small businesses to submit very complicated forms at frequent intervals. Gerry can't fulfil his current tasks as well as fill out the forms, so he needs Sandra to take over some of his jobs. She could either do the tailoring or deal with the customers, but since Gerry often feels awkward with customers he would prefer to delegate this to Sandra, even though it isn't what he originally hired her for. Sandra prefers the tailoring, and is better at it than customer relations, but can do the latter well enough and accepts the new task.

The issue here is that in order to be viable Gerry's business must satisfy a complex set of constraints, including doing the jobs that bring in paying customers and satisfying the obligations imposed by the taxation authorities. Under a particular set of conditions, the business can settle into a specific functional task distribution that satisfies those constraints. This occurred during the process in which Gerry hired Sandra and they adjusted the functioning of the business based on her original job description. However, an unexpected perturbation can change the viability constraints on the business and make the old task distribution unworkable. What matters at this point is whether the business can redistribute the functional load in a way that satisfies the new constraints. Gerry no longer wants Sandra simply to fulfil her original job description, he wants her to take on new tasks. There may actually be several kinds of redistribution that will work, with perhaps only minor differences in relative advantage between them. The most important thing is that any functional redistribution that occurs must maintain the viability of the business. This is the basic norm applying to Gerry's management problem. For this reason, *the norms of the business shouldn't be uniquely associated with a specific set of tasks.*

This holistic structure is a general feature of the norms that apply to intelligent agents. Agents need to remain viable through a coherent pattern of activities, or lifestyle (broadly construed). When a problem arises, the agent must identify the problem and attempt to determine its

ramifications for the agent's overall activities. The agent then needs to perform compensatory action that re-establishes coherency in the activity complex. This might involve a minor change in activity that 'tweaks' an existing lifestyle, or it might involve a major shift in the lifestyle itself. Gerry's brother Frank was forced to give up his plans to become a dentist when he discovered that he found looking into people's mouths revolting. He became an accountant instead.

Thus, performance envelopes determine norms for an agent, and it is an important characteristic of norms that there are typically many of them and they act in concert. For this reason we describe the norms that apply to adaptive agents in terms of a norm matrix rather than in terms of individual goals. This makes an important difference, because when faced with multiple norms an agent must find the best balance between them. For cognitive agents in realistic contexts this is the central, not the derivative, decision-making situation.

There is a further distinction that is important for understanding norms and agency, namely the distinction between **implicit** and **explicit** norms. We can illustrate this difference in terms of what Gerry does and does not know about his management problems. There are many features of the operation of Gerry's business that he doesn't understand and consequently is unable to fix if they go wrong. For instance, Gerry doesn't realise that his personality grates on Sandra, and that as a result she is working much less hard than she could. Gerry fails to detect this problem both because he isn't very good at reading Sandra's state of mind, and also because he doesn't have any previous experience with hired staff on which to base expectations of productivity. On the other hand there are other kinds of problems that Gerry can detect: Gerry can tell when the books don't balance, he knows when jobs aren't completed on time and when customers complain, and he knows that if he fouls up his tax forms he'll be audited. Consequently, in analysing the viability of Gerry's business and the nature of his management problems, we can make a distinction between norms that Gerry can identify and those he can't. We shall refer to the norms that Gerry can identify as his **explicit norm matrix**.

The explicit norm matrix is of enormous significance because it provides the steering information for Gerry's management decisions. Gerry's ability to keep his business viable depends on whether his explicit norms (balancing the books, completing jobs on time, etc.) guide his actions well enough that they sustain the overall viability conditions of the business. If problems develop that Gerry can't recognise or take action to correct then the business may cease to be viable.

The distinction between implicit and explicit norms also applies in non-human contexts. Affective processes (aversion and reward) provide one of the fundamental steering mechanisms for organisms, and we shall refer to the array of affective processes that an organism possesses as its explicit norm matrix. Thus, hunger indicates that the animal's feeding requirements are not currently being met, while satiation indicates that at the moment they *have* been met. In this respect, affect processes are explicit, not necessarily because they are conscious, but because they provide a direct informational pathway for evaluating action. They do this by forming a direct part of the neuro-anatomical control of motor action. On the other hand, the underlying systemic conditions to which they correspond, e.g. inadequate cellular nutrition in the case of the hunger signal, are typically not explicit for organisms; it takes a science of nutrition to uncover

the details of the nutritional requirements that underlie hunger.

By providing explicit norms, affective processes have wide-ranging effects on adaptive interaction capacity. They permit behavioural flexibility by specifying goals for action rather than specifying action directly (cf. Rolls 2000). They also serve as a mechanism for integrating multiple factors in action production because many affective norms can apply to a given action, and properties such as relative intensities allow comparison and trade-off amongst affective signals. For example an animal might use relative intensity of thirst and hunger signals to determine whether it seeks water or food in a particular context.¹⁰ Furthermore, affective norms allow organisms to learn about their implicit norms through processes such as **stimulus reinforcement association** (Rolls 2000) and **predictive reward learning** (Montague and Sejnowski 1994). Essentially these learning processes work by allowing an organism to associate aspects of interaction with reward information, which is a fundamental requirement for skill construction. Thus, norms play a fundamental role in interaction ability and the processes of cognitive development.

3 At the threshold of self-directedness

3.1 *Of mosquitoes, bumblebees and cheetahs*

The concept of self-directedness is designed to capture the distinction between reactive and anticipative forms of adaptiveness. Self-directedness involves the ability to acquire information from interaction and to use it to modify performance so as to satisfy the agent's norms. To convey a clearer sense of what we mean by self-directedness we first contrast mosquito blood-host search behaviour, which by our account is not self-directed, with cheetah hunting, which is relatively strongly self-directed. We then turn to examining bumblebee foraging, which lies right at the threshold of self-directedness.

Mosquitoes are morphologically relatively simple, and depend on a comparatively simple set of niche relations to complete their life-cycle. One of the most important requirements of this life-cycle is that females acquire blood in order to produce eggs. Females locate blood hosts by locating chemicals, including carbon dioxide, produced by blood hosts. The simple chemotactic process of flying in the direction of increasing carbon dioxide concentration brings a mosquito into proximity with a blood host, whereupon feeding behaviour is initiated (Klowden 1995).

Cheetahs, on the other hand, are morphologically more complex and in particular have much greater nutritional requirements and much more complex sensori-motor systems than mosquitoes. The niche relations they exploit are correspondingly more complicated, and most significantly involve a number of variables to which mosquitoes are insensitive. For a mosquito, blood hosts are common and indistinguishable; any blood host will do. In contrast, cheetahs must be highly sensitive to both prey type and context. Large and dangerous animals can injure them, they can expend too much energy trying to catch fast healthy animals, and different species and different individuals have different flight/fight strategies, etc. For these kinds of reasons there are no simple reliable signals that indicate suitable prey, comparable to the role carbon dioxide plays for

¹⁰ Christensen and Hooker 2000, Raubenheimer and Bernays 1993, Rolls 2000.

mosquitoes. Cheetahs must learn to recognise appropriate prey using complex, context-sensitive discrimination honed by experience. Moreover, simply travelling in the direction of the prey is unlikely to result in catching it. Cheetahs must tailor their actions to the behaviour of the prey by stalking it and responding to its movements during the chase.

With respect to understanding self-directedness there are several noteworthy features of this contrast. Most obviously, although both species are adapted – mosquitoes more widely so than cheetahs – cheetahs have a greatly elaborated ability to shape their actions to the environmental context. Furthermore, achieving this context-sensitivity crucially involves the ability to coordinate many factors, simultaneously and over time. Whereas mosquito behaviour has a highly modularised organisation in which each type of action, such as carbon dioxide tracking, is governed by at most a few signals, cheetah behaviour is highly integrative; many kinds of signals are used to shape action at any given time, and the response to particular types of stimuli is context sensitive. For instance, when it is extremely hungry, a cheetah may attempt to catch types of prey that it would ignore if it were less hungry. These processes of integration play a key role in the context-sensitivity of cheetah behaviour, both because they can allow a given action to be shaped by many sources of information, and because they permit the propagation of information to many relevant activities. Learning is an important part of this. The sheer number of interrelated factors involved in successful hunting, such as available cover, stalking distance, prey speed and agility, means that cheetahs must learn many of the relevant relationships through experience. For instance, learning to stalk to a sufficiently close distance depends on discovering the relative speed and agility of the prey, and its characteristic sensory acuity.

Mosquitoes are not self-directed on our account because they don't anticipate interaction processes, and hence cannot modify their responses context sensitively. Instead, they react to local stimuli with fixed behaviours. On the teleosemantic theory of intentionality mosquitoes *do* anticipate, since the meaning of the CO₂ concentration signal is interpreted as something like 'a blood host in this direction' (cf. Millikan 1989, 1993). This interpretation gives the impression that mosquitoes anticipate that by flying up a CO₂ gradient they will arrive at a blood host. Now, there is a certain respect in which this interpretation makes sense. Because CO₂ gradients often enough culminate in a blood host, following them is an adaptively successful behaviour for mosquitoes. However, there are other important respects in which it is highly misleading to interpret mosquitoes as anticipating that a CO₂ stream will culminate in a blood host.

One important problem is that there is no informational pathway active in the control of flight behaviour that associates CO₂ concentration with arriving at blood hosts. The blood search process is organised in terms of serial action modules: CO₂ governs a particular parameter of the operation of a particular behaviour module, namely the spatial orientation of flight. Proximity to a blood host engages a separate feeding behaviour module. There is no process in the flight module that connects these relationships. The connection is made through the environment, which scaffolds the overall organisation of the interaction process, not by motor control processes internal to the mosquito. In other words, there is nothing in the architecture of the CO₂-tracking module that primes it for culmination in proximity to a blood host as a specific event amongst a variety of kinds of outcome that can occur. In this respect, no recognition of the outcome is involved in the control of the action, and there is no learning about outcomes. Thus, on the not unreasonable proposal that anticipation involves some form of expectancy derived

from experience, mosquitoes do not anticipate arriving at blood hosts. To be sure, it is an implicit normative requirement (in the sense we characterised in §2.3) of the overall process organisation of the mosquito life cycle that CO₂-tracking results in proximity to blood hosts. Nevertheless, the relationship between action and outcome is not explicitly differentiated by mosquitoes in the control of action.

One way to pose this distinction is by looking at the processes by which the relations between flight control and blood host location can be modified; specifically, either by mutation and preferential selection or by external intervention in design. The key point is that the *mosquito itself* cannot modify the relation.

Bumblebee flower foraging provides a biological example of how on-board modification of action-outcome relations can occur. It is an example of a behavioural process that has a level of complexity close to that of mosquito blood-host search, but with the important difference that anticipations about the outcome of action do play a role in motor control, making bumblebees minimally self-directed. Mosquitoes don't need to explicitly anticipate the outcome of CO₂-guided flight behaviour because the adaptively significant relationship between CO₂ and blood hosts is stable. Populational adaptation has been sufficient to find and sustain a simple correlation that is sufficiently adaptive. For bumblebees, however, the relationships between the availability of flower types, flower colour, and nectar yields are both variable (over species, times, and locations) and adaptively significant. In this case populational adaptation would be unable to find the appropriate correlations rapidly enough for adaptive success because the relations are spatially and temporally variable relative to the life cycle of bumblebees. Differentiation of the correlations must be performed by a process that is more rapid than the rate of change of the correlation, and is complex enough to carry out the cross-correlational signal processing required to extract the relevant adaptive relationships. Bumblebees solve the problem by learning when foraging. They sample the flower types within their range, then preferentially visit those with an adequate nectar reward (Real 1991).

In so doing, bumblebees illustrate a simple ability to interactively differentiate an adaptively relevant relationship and use it to modify behaviour. It is worth identifying some of the capacities that underlie this ability (see Montague et al. 1995). Bumblebees possess:

- An ability to differentiate environmental stimuli (colour of flowers, amount of nectar).
- An ability to differentiate the affective value of interactive relations (a preference for greater nectar quantities, an ability to associate nectar quantity with flower colour).
- An ability to modify behaviour (modify the type of flower visited).

What makes bee foraging behaviour self-directed is the connection between the affective evaluation of the outcome of flower visits and the modification of subsequent flower visitations. The bees are capable of a very simple form of learning, specifically, a very simple form of anticipation, constituted by the bias to fly towards flowers of a particular colour. Thus, after learning, there is a significant sense in which *the bee itself* anticipates that visiting a flower of a particular colour will result in reward.

3.2 Some implications for intentionality

Self-directedness depends on utilising and cross-correlating information. It therefore presupposes *some* form of intentional content. However, self-directedness also imposes some constraints on the nature of intentional content. In particular, such content must be interpretable by the system; the system must be able to evaluate the information and relate it to other information, including by cross correlation. This is a constraint that the teleosemantic theory of intentional content doesn't satisfy, because teleosemantic content is specified in terms of the organism's selection history, and organisms typically have no access to information about their selection history. So organisms have no way of evaluating or cross-correlating teleosemantic content.

In this respect it is significant to note that the form of anticipation characterised above satisfies the criterion for misrepresentation that preoccupies teleosemantics. A bee's anticipations can be wrong, since the particular flower visited may not have nectar, or not enough nectar. Moreover, not only can they be wrong, the bumblebee can detect the error in the form of a reduced gustatory reward from the flower. So the anticipation is a form of information which the organism itself can evaluate.¹¹

It would take us beyond our current focus to develop a detailed theory of intentional semantics, however several features relevant to such an account suggest themselves. The fundamental adaptive problem that signal utilisation solves is the control of the nature and timing of actions an organism performs. From this perspective the most natural way to interpret the information a signal provides for an organism is in terms of the difference the signal makes to the actions the organism performs. In the basic case, then, the norm that applies to information utilisation is, 'is the action performed successful?', rather than, 'does the signal correspond to the appropriate object or external state of affairs?' There are general adaptive reasons for preferring this interpretation, since the success of action is more significant than accurately representing objects or states of affairs. Moreover, an action's success is something that organisms can and do evaluate readily, whereas evaluating correspondence between representation and represented is a task that is, at best, complex and resource intensive, frequently impractical and, especially among simpler agents, often impossible.¹² The success and failure of action is therefore likely to play a more basic role in learning and cognition than is reference. Relating the fundamental form of information to action gives information, partly via affect, a common currency that the system can use as a means for relating multiple sources of information. This makes it information 'from the system's perspective', and thereby likely to be the basic form of information relevant to cognitive processes.

To place the issue of cognitive relevance in perspective it is worth contrasting the standards for attributing intentional content employed by teleosemantics with standards used in developmental psychology. As articulated by Millikan, the *raison d'être* of teleosemantics is explaining

¹¹ See Bickhard 1993 for a theory of representation based on indication and system-detectable error.

¹² Insofar as the represented is taken to be the source of the signal, it is physically impossible, since the source is in the past. It is usually assumed that the represented is a temporally persistent object, but this will often not be the case, especially amongst simpler organisms. We thank Mark Bickhard for pointing this out.

misrepresentation.¹³ For example, in the case of the frog tongue-flicking behaviour the content of the perceptual small-dark-moving-object stimulus is supposed to be ‘bug here now’. The fact that frogs will also flick their tongues at bee-bee pellets can thereby be explained as a misrepresentation, since in this case the stimulus doesn’t correspond to the type of object that made the behaviour adaptive and resulted in its selection. But compare this attribution with a situation in which a psychologist is attempting to determine the conceptual knowledge of a young child. The psychologist shows the child a picture of a horse and a truck, and says, ‘which one is the horse?’ Even if the child points to the horse, there is not yet enough evidence to justify assuming that a picture of a horse means ‘horse’ to the child. The psychologist next shows the child a picture of a horse, a cow, a pig and a dog, and again asks to be shown the horse. This time, though, the child is uncertain. When the question is repeated the child points at the dog. Now the evidence points in the other direction, suggesting that the child doesn’t properly understand the concept of ‘horse’. To ensure that this isn’t a one-off error the psychologist repeats the experiment a number of times, each time using different pictures of the same animal types to control for the possibility that some feature of the pictures is confusing the child. If the child tends to get it right most of the time then the psychologist is likely to interpret this as meaning that the child does in fact understand the concept of horse, but makes occasional errors. On the other hand, if the child makes persistent errors then the psychologist will take this as evidence that the child doesn’t understand what ‘horse’ means. The child can differentiate four-legged animals from vehicles, and associate the word ‘horse’ with the animals, but she can’t be more specific than this.

Based on this kind of experimental methodology, a psychologist would not attribute the representation ‘bug here now’ to a frog. Even though the frog gets it right on one kind of discrimination task, the fact that the frog persistently gets it wrong on another similar task would be sufficient for the psychologist to decide that the frog doesn’t really understand the concept of ‘bug’. We believe that the psychologist’s interpretation is preferable to the teleosemantic one because it makes assumptions about the nature of intentional content that are more stringent and better attuned to cognitive relevance. Specifically, the psychologist’s experimental methodology assumes the following:

- Intentional content should be attributed based on the actual discriminatory abilities of the subject. If the representation is supposed to be of an object type, the subject should be able to robustly identify the object type under a range of conditions, including against a range of relevant contrasts.¹⁴
- Consequently, it is important not to attribute more features to a concept held by the subject than the subject can reliably differentiate. E.g. it might be argued that the child has a relatively undifferentiated concept of ‘animal’ with which she associates the word ‘horse’, but not a concept of ‘horse’ per se.

¹³ See especially Millikan 1993, introduction.

¹⁴ Deciding what counts as a relevant contrast is clearly crucial, but there is no a priori answer. Which alternatives are relevant depends on the concept and the context of use. Clearly concept possession cannot be required to rule out all logically possible contrasts, but equally clearly it should rule out some contrasts in practice.

- Occasional error may be the result of misrepresentation¹⁵, however persistent error is evidence that the subject doesn't have representational competency.

The problem with the teleosemantic interpretation of content is that it is unduly generous. In the frog case teleosemantics attributes representation of an object type without requiring that the subject be able to differentiate distinctive attributes of the object type or differentiate the object type from contrasting object types. In this respect it is interesting to note that, using the looking time methodology, Xu and Carey (1996) find that infants do not track the identity of objects by type until after 10 months of age.¹⁶

With respect to self-directedness the most important fundamental issue concerns error. If intentional content is to be in a form that is interpretable by the agent, then the agent must be able to detect when the content is wrong. As we noted, the teleosemantic account cannot satisfy this criterion, and its major claim to fame is that it solves the problem of the possibility of misrepresentation! To appreciate the implications of this issue it is helpful to draw a comparison with the internalist/externalist debate in epistemology.¹⁷ Externalists claim that the justification for belief is concerned with the nature of the process connecting the belief to its referent (which need not be understood by the believer), whilst internalists require that the believer should understand or have access to the warrant for belief. Although internalism is regarded by some of its adherents as involving a stand against naturalist epistemology¹⁸, Kitcher (1992) points out that naturalist approaches need to characterise real world knowledge processes as self-correcting in order to retain normative ambitions. But this means that epistemic norms must be accessible to epistemic agents - at least in part - otherwise there could be no *self*-correction. The same reasoning applies to naturalist theories of representation: the ability to learn about representational content is an essential postulate for an adequate naturalist theory of representation, and the ability to detect misrepresentation is required in order to learn about content. Teleosemantics can provide no role for misrepresentation in cognition, and this is a serious problem. It should be noted that the detectability-of-error criterion for content that we are proposing does not imply that agents cannot misrepresent, it implies that if they do misrepresent then they should be able to discover the fact. This is why persistent error is evidence, not of misrepresentation, but of failure to represent.

4 Improving self-directedness

In this section we will look at how self-directedness improves, and some of the implications this has for intentionality and high order cognition. In essence, self-directedness increases in strength as the processes for targeting action become more sophisticated in the way they coordinate the

¹⁵ Or carelessness, or a misunderstanding of the question, etc.; what is important, however, is that they all tend to be occasional, rather than consistent over time.

¹⁶ This and related experiments are discussed in Hauser and Carey 1998.

¹⁷ We thank Jill McIntosh for drawing this to our attention.

¹⁸ E.g. Fumerton 1988.

interaction process. As organisms become increasingly self-directed they are better able to manage complex variable interaction processes, and begin to exhibit distinctively cognitive processes such as choice and planning. This primarily occurs through increases in the ability to anticipate and evaluate. Bees are only weakly self-directed because their capacity to form anticipations is fairly limited. Stronger forms of self-directedness require more powerful learning processes for forming anticipations.

Some kinds of interaction processes show nonlinear sensitivities, in that variation in any of a number of factors may produce highly divergent interaction pathways, many of which have effects that are not adaptive for the system. For example, a small mistake when a cheetah is stalking a gazelle can alert the gazelle and allow it to escape. Tasks of this nature impose particularly strong demands on an organism's capacity for anticipation, since the organism must initiate and be responsive to extended temporal patterns in interaction that involve many interdependent factors.

Improvements in anticipation capacity allow the system to shape its actions over longer timescales and with respect to more detailed, in some cases modal, information concerning the interaction process.¹⁹ As we saw in the bee example, evaluation plays an important role in the process. Evaluative signals function to cross-correlate the control of action with the success of its outcome.

A powerful interactive learning process called **Self-Directed Anticipative Learning (SDAL)** can be generated by the coupling of anticipative and evaluative processes. SDAL is effective for solving open problems, in which the nature of the task to be performed is not known in advance. SDAL uses interaction to acquire information about the nature of the task and thereby improves performance. The system learns from experience and modifies its behaviour, continually tracking the success of subsequent modifications. As the system interacts it generates information that allows it to construct anticipative models of the interaction process; in turn these anticipations modify interaction, which allows the system to perform more focussed activity and generates further feedback to the system. This feedback serves to evaluate the success of the anticipations, whilst the anticipations themselves help the system improve its recognition of relevant information and evaluate its performance more precisely. If the anticipations prove unsuccessful, the system will hunt fruitlessly, but if they are even partially successful the system can progressively improve its ability, bootstrapping its way to a solution.

An example of SDAL is the process by which cheetahs learn to hunt. Gaining the skills required for successful hunting requires extensive learning, in which cheetahs evaluate their own performance and use information from interaction to improve performance. As cubs, cheetahs spend a great deal of time learning hunting skills by playing with siblings, chasing lizards, and so forth. The mother facilitates this process by bringing small live prey, such as a hare, back to the cubs, allowing them to practice chasing and killing techniques. As the cubs begin to mature they accompany the mother on hunts and observe the real process first-hand. Even so, actual hunting experience is required before proficiency is achieved; many juveniles, for instance, make the

¹⁹ See further Christensen and Hooker 2000. For some of the neural mechanisms involved see Montague and Sejnowski 1994. See also Glenberg 1997 for an interactivist interpretation of the role of memory.

mistake of initiating the chase from too great a distance. The hunting capacity of a mature cheetah is thus a complex product of an extended history of mutual shaping between internally generated action and the success and failure of the ensuing interaction processes.

Hunting is an integration problem: prey is caught only when a complex array of factors are brought into coordination. Achieving this coordination requires detailed knowledge. To hunt successfully cheetahs must differentiate many specific objects (such as the prey and obstacles) and relations (such as distance to the prey and to flight/fight opportunities). Part of this differentiation process involves recognising sources of error, such as startling the prey too early, or tackling an animal that is too large. These differentiations are not simply given to a cheetah perceptually, they are acquired through an extended interactive learning process, and they concern interactive characteristics (alertness, speed, agility, aggressiveness, etc., relative to the cheetah's behaviour).

On our approach to intentionality, cognitive reference arises through these interactive differentiation processes. Specifically, as part of the problem of coordinating complex interaction processes, self-directed agents learn to differentiate specific states-of-affairs, objects, and object types. Our hypothesis is that they do this by learning the effects these things have on interaction processes. In this context it is worth briefly discussing a standard argument against associating content with action, which is that representations might be used to indicate an open-ended range of actions, depending on the circumstances. For example, representation of the presence a chair might lead an agent to sit on it, stand on it to change a light bulb, use it to block the door, belt an intruder over the head with it, break it up for firewood, and so on.²⁰ This observation is supposed to justify an in-principle separation of action and content, the idea being that, because representations need not be associated with any specific actions, their content isn't connected to action at all. But the argument hardly justifies this conclusion. At most it shows that some representations are relatively open with respect to action possibilities. But this does not mean that they are disconnected from action. Cognitive agents learn to represent *through* interaction, and even sophisticated representations are grounded in interaction relations.

The capacity to represent multiple action possibilities can be understood from an interactionist standpoint by examining the nature of the developmental learning processes that give rise to representation and concept formation. These learning processes involve, in part, the agent partitioning interaction processes into increasingly finely differentiated categories by learning to recognise important sources of influence on the nature of the interaction. A fundamental distinction to be made concerns effects that arise from the agent versus effects that arise from elsewhere. It is to be expected that evolution will impose biases on the developmental processes of cognitive agents that facilitate differentiating objects, and important object and event types, but this should not be interpreted as evolution imbuing innate knowledge of those things (see Karmiloff-Smith 1992, who argues for this type of interpretation of cognitive development). Nor will cognitive agents suddenly come to acquire concepts of event and object types as they are encountered, or at some specific later point. For instance, a child's concept of apples is acquired progressively as the child learns what an apple looks like, feels like, tastes like, etc. A child does this by interacting with apples, and as the range of interactive experiences increases, the child's

²⁰ Cf. Millikan 1989, p.289-90.

concept of an apple gains increasing richness. Even after the stage when a child is able to use the concept of apple appropriately in many circumstances, such as in normal conversation and at lunchtime, the child may still be learning about the nature of apples. It comes as a pleasant surprise to many children that apples explode nicely when hurled at brick walls.

A key feature of this learning process is a progressive shift from crude to increasingly fine differentiation of interaction characteristics, linked to the child's improving sensorimotor skills. As learning progresses the child will gain enough information about the interaction characteristics of apples to spontaneously associate apples with novel actions within a particular range of action types, such as being able to know without specific experience that one can juggle apples. Consequently the concept becomes less tied to specific action contexts.

However, this doesn't mean that the child's representation of apples is fundamentally separated from action, it just means that the child knows enough about the interaction characteristics of apples to relate apples to a particular range of sensorimotor skills. In other words, once the child knows what an apple feels like to grasp and throw, the child can generalise by using apples in new actions within the child's range of grasping and throwing abilities. Nonetheless, the action range within which a child can use a concept will show a high level of experience-dependency, notwithstanding the fact that the child is able to generalise to novel actions *within* the range. Thus, although a young boy may, with no prompting or prior experience, throw an apple at a window in order to break it, that same boy is extremely unlikely to know how to prepare a pork and apple pie.²¹

Thus, the argument for divorcing content from action has the situation on its head. It is a very important feature of representations that they can sometimes be used to indicate open-ended action possibilities. But this doesn't justify a fundamental, in-principle separation of action and content. Our way of representing the world is experience-dependent. And, concept boundaries – the ways in which an agent decides that something *doesn't* belong to a category – are also strongly related to action. If a person sees a chair and attempts to sit on it, but falls through thin air onto the floor, the person is likely to decide that what they see isn't really a chair. It might be a hallucination, or a hologram. Similarly, if a child tries to bite into an apple and gets a mouth full of wax, the child will probably decide that the object isn't really an apple.

There are important reasons for preferring an interactivist account of reference of this type to standard teleosemantic ones. We discussed one of the most fundamental earlier; interpreting information in terms of action acknowledges and incorporates the fact that agents interpret information and utilise it flexibly. Also, there is extensive empirical evidence that human concepts have a highly interaction-oriented character. According to prototype theory, categories are grounded in interaction properties as experienced from the perspective of the cognitive

²¹ To put this in the terms we introduced earlier, concepts are closely associated with the performance envelopes and norm matrices of the agent that apply in the interaction context. Young boys have good performance envelopes for throwing, and consequently their concept of apple can include a rich understanding of throwing potential. In contrast they generally have poor cooking performance envelopes, and so little understanding of apples in relation to cooking potential. Moreover, since they have a very limited ability to tell good cooking from bad, their ability to learn about the cooking potential of apples is similarly restricted. Even though they might memorise recipes involving apples, they won't be able to exercise judgement or improvise in the way that a skilled chef can.

agent.²² Furthermore, the interactivist account fits well with evidence in developmental psychology concerning the dynamical nature of cognitive development and the importance of interaction²³, and with evidence in neuroscience concerning massive levels of activity-dependency in neuronal organisation, and consequently very high-levels of experience-dependency in neuronal functional organisation.²⁴

For all the emphasis placed on representation and categorisation by most approaches, the integrative aspects of intentionality are no less important. It is a mistake to assume that cognitive integration processes occur as operations on atomistic representations; for the reasons discussed in §2.3 and §3, integration is a more basic process than representation and in fact takes priority in interaction and learning. Moreover, integration is a feature of high level cognition as exemplified in the phenomenon of holistic situational awareness in highly skilled activities.²⁵ Playing professional tennis, for example, is a highly cognitively demanding task that requires considerable strategic skills and acute situational awareness. A player with clear physical advantages, such as more powerful groundstrokes, can be outplayed by a player who is able to shape the game so that those strengths are negated. One of the most cognitively demanding aspects of professional tennis is that it requires highly developed anticipations concerning the performance interrelationships of the game. Some of the kinds of things professional tennis players must be able to anticipate include the differing characteristics of a baseline player as opposed to a serve-volleyer, and the effects that playing on a grass or a rebound-ace or a clay surface has on each style of play. The only way to acquire these anticipations, at least to a level sufficient for being competitive, is through a long learning process involving experience of actual game conditions.

The role of these anticipations is to focus action appropriately for the variety of conditions the player will experience, and to facilitate the rapid localisation of success and error needed to adapt effectively within a match. Note that although a novice player may have a conscious goal to win a game, this goal has little anticipative content since the novice has no understanding of the kinds of performance relations it involves. As a result, the novice flounders in a morass of unfamiliar relations that must be somehow coordinated to play well. The unfamiliarity of the situation means that the novice has little ability to recognise sources of error or to take well-directed corrective measures, i.e. she cannot differentiate the adaptively relevant relations for the context. In contrast, the professional player's performance anticipations, built up through years of

²² These include prototype effects themselves, namely the fact that often some members of categories are treated as more typical than others, where what determines typicality is some aspect of the agent's physical makeup or interaction experience, such as treating primary colours as more typical than non-primary colours, or treating robins as more typical examples of birds than penguins (Lakoff 1987, Rosch 1973). They also include the phenomenon of basic level categories, which, roughly, are categories most commonly used, have the simplest names, are the first to be learned, and are distinguished by commonly experienced interactive attributes (Lakoff 1987, 46-7). Examples of basic level categories include *dog*, *chair*, *book* and *car*.

²³ E.g. Glenberg 1997, Karmiloff-Smith 1992, Smith and Thelen 1993, Thelen 1995.

²⁴ E.g. Christensen and Hooker 2000, Florence et al. 1998, Jones and Pons 1998, Quartz and Sejnowski 1997.

²⁵ Cf. Merleau-Ponty's (1962) account of intentionality. See also Dreyfus 1996 and Brown 1988.

coaching and match play, allow her to learn quickly about the opponent's characteristics and match her performance to their strengths and weaknesses. This may be as detailed as being able to anticipate the likely direction of the serve at set point when the score is 30-40, or anticipating that the opponent's service game will be likely to crack under the pressure of a tie-break. The professional's anticipations help localise success and error by making salient important game relations. For instance, if the opponent is attempting to disrupt the player's baseline strategy by frequently coming to the net, the player may counter by attempting some dramatic low percentage passing shots which, if successful, may reduce the other players confidence and create doubt about when to go the net. {perhaps you could put this (in some truncated version) in a footnote. Although it helps make your point well, I don't think it warrants the space in the main text. I'm very reluctant to cut this, and Cliff has said the same. The problem is that we are challenging a tradition that fails to recognise this kind of phenomena as being a part of sophisticated cognition. We think that the existence of these phenomena provides some of the strongest support for our approach. The rationale for the conceptual work is much harder to understand if people don't have a clear picture of the kinds of things we're trying to explain. If space really does have to be made I'll see what I can do to trim things, but it is only a paragraph and a half.

5 Conclusion

This account of self-directed agents develops a perspective for drawing together diverse issues involved in the evolution of cognitive agency, including intentionality. Here is a summary of the main points of our account:

- **Autonomy:** Adaptive systems are composed of networks of processes that are interdependent and collectively self-sustaining and self-repairing. The theory of autonomy is an analysis of the identity conditions for such systems. The concept of performance envelopes provides a way of understanding the adaptive behaviour of autonomous systems from an open-ended dynamical perspective rather than in terms of fixed task specifications.
- **Norm matrices:** This in turn provides us with a radically different conception of norms to etiological theory. The concept of a norm matrix identifies norms with conditions of viability, and provides a way of characterising the fact that in realistic biological conditions many norms are operative simultaneously. Adaptive interaction requires the continuous satisfaction of many norms. An explicit norm matrix is the array of normative conditions that an agent can explicitly recognise. In biological organisms, affect processes provide the basic explicit norm matrix.
- **Anticipation and evaluation:** Explicit norm matrices provide steering information for behaviour, and in particular provide a basis for 'on-board' processes that modify action-outcome relations. Evaluation allows an agent to assign an affective value to features of interaction to modify performance accordingly, and to anticipate future outcomes.
- **Interactive differentiation:** Organisms use signals to control the nature and timing of the actions they perform. More complex intentional content arises from the information processing involved in learning to recognise the various contributing factors to interaction, tracking sources of success and error.
- **Self-directedness:** Anticipation and evaluation combine to generate the capacity for self-directedness. Self-directedness involves the ability to acquire information from

interaction and use it to modify performance so as to satisfy the agent's norms. As agents become increasingly self-directed they are better able to manage complex variable interaction processes, and begin to exhibit distinctively cognitive processes such as choice and planning.

- **Self-directed anticipative learning:** In certain circumstances anticipation and evaluation can be mutually amplifying, as more focussed action and improved interactive differentiation further improve anticipation and evaluation. We hypothesise that these powerful learning processes play a central role in cognitive development. We further think it is likely that reference and concept formation occurs through these processes as agents come to differentiate object and event types in interaction.

Rather than attempting a simplistic unification of representationalist cognitive science and adaptationist evolution theory, our account focuses on important adaptive issues that are likely to have played a role in shaping the phylogeny of intelligence. It develops an account of intentional agency that is grounded in biologically realistic adaptive problems, and which coheres well with a number of strands of research in contemporary cognitive science, including autonomous agent robotics, developmental psychology, cognitive psychology, and cognitive neuroscience.

Bibliography

- Beer, R.D. (1995) "Computational and dynamical languages for autonomous agents", in *Mind as Motion: Explorations in the Dynamics of Cognition*. R.F. Port, T. van Gelder (eds) Cambridge, MA: MIT Press.
- (2000) "Dynamical Approaches to Cognitive Science", *Trends in Cognitive Sciences*. Vol 4, pp. 91-99.
- Bickhard, M.H. (1993) "Representational content in humans and machines", *Experimental and Theoretical Artificial Intelligence*. Vol. 5, pp. 285-333.
- Bickhard, M. H., and Ritchie, D. M. (1983) *On the nature of representation: A case study of James J. Gibson's theory of perception*. New York: Praeger.
- Brooks, R.A. (1991) "Intelligence without representation", *Artificial Intelligence*. Vol. 47, pp. 139-159.
- Brown, H. I. (1988) *Rationality*. London: Routledge.
- Cherniak, C. (1986) *Minimal rationality*. Cambridge, Mass.: Bradford/MIT Press.
- Christensen, W.D. and Bickhard, M.H. (To appear) "The Process Dynamics of Normative Function", *Monist*.
- Christensen, W.D., and Hooker, C.A. (2000) "An interactivist-constructivist approach to intelligence: self-directed anticipative learning", *Philosophical Psychology*. Vol 13(1), pp. 5-45.
- (2000) "Autonomy and the emergence of intelligence: Organised interactive construction", *Communication and Cognition - Artificial Intelligence* 17(3-4): 133-157.
- (To appear) "Representation and the Meaning of Life", in *Representation in Mind: New Approaches to Mental Representation*. H. Clapin, P. Slezak and P. Staines (eds), Westport: Praeger.
- Clark, A. (1997) *Being there: putting brain, body, and world together again*. Boston: Bradford/MIT Press.
- Dreyfus, H. (1996) "The current relevance of Merleau-Ponty's phenomenology of embodiment",

- in *Perspectives on embodiment*. H. Haber and G. Weiss (eds), London: Routledge.
- Eaton, R. L. (1974) *The cheetah; the biology, ecology, and behavior of an endangered species*. New York: Van Nostrand Reinhold Co.
- Florence, S., Taub, H. and Kaas, J. (1998) "Large scale sprouting of cortical connections after peripheral injury in adult macaque monkeys", *Science*. Vol. 282, pp. 1117-21.
- Fumerton, R. (1988) "The internalism/externalism controversy", *Philosophical Perspectives*. Vol. 2, pp. 443-459.
- Glenberg, A.M. (1997) "What memory is for", *Behavioral and Brain Sciences*. Vol. 20, pp. 1-55.
- Griffiths, P. (1997) *What emotions really are*. Chicago: University of Chicago Press.
- Griffiths, P. and Gray, R. (1994) "Developmental Systems and Evolutionary Explanation", *Journal of Philosophy*. Vol. 91, pp. 277-304.
- Hauser, M. and Carey, S. (1998) "Building a cognitive creature from a set of primitives: evolutionary and developmental insights." In *The Evolution of Mind*, D. Cummins and C. Allen (eds). NY: Oxford University Press.
- Hendriks-Jansen, H. (1996) *Catching Ourselves in the Act: Situated Activity, Interactive Emergence, Evolution and Human Thought*. Cambridge, MA: MIT Press.
- Jablonka, E. and Lamb, M.J. (1995) *Epigenetic inheritance and evolution: the Lamarckian dimension*. Oxford: Oxford University Press.
- Jones, E. and Pons, T. (1998) "Thalamic and brainstem contributions to large-scale plasticity of primate somatosensory cortex", *Science*. Vol. 282, pp. 1121-25.
- Karmiloff-Smith, A. (1992) *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.
- Kauffman, S.A. (1993) *The origins of order: self-organization and selection in evolution*. New York: Oxford University Press.
- Kitcher, P. (1992) "The naturalists return", *Philosophical Review*. Vol. 101, pp. 53-114.
- Klowden, M. J. (1995) "Blood, sex, and the mosquito: control mechanisms of mosquito blood-feeding behavior", *BioScience*. Vol. 45, pp. 326-31.
- Lakoff, G. (1987) *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*. Chicago: The University of Chicago Press.
- Merleau-Ponty, M. (1962) *Phenomenology of perception*. Transl. C. Smith. London: Routledge & Kegan-Paul.
- Millikan, R.G. (1989) "Biosemantics", *The Journal of Philosophy*. Vol. 86, pp. 281-97.
- (1993) *White Queen Psychology and Other Essays for Alice*. Cambridge, MA.: MIT Press.
- Montague, P.R., Sejnowski, T.J. (1994) "The Predictive Brain: Temporal Coincidence and Temporal Order in Synaptic Learning Mechanisms", *Learning & Memory*. Vol. 1, pp. 1-33.
- Montague, P.R., Dayan, P., Person, C. and Sejnowski, T.J. (1995) "Bee foraging in uncertain environments using predictive hebbian learning", *Nature*. Vol. 377, pp. 725-8.
- Oyama, S. (1985) *The Ontogeny of Information*. Cambridge: Cambridge University Press.
- Pfeifer, R., and Scheier, C. (1999) *Understanding Intelligence*. Cambridge, M.A.: The MIT Press.
- Quartz, S.R., & Sejnowski, T.J. (1997) "The Neural Basis of Cognitive Development: A Constructivist Manifesto", *Behavioural and Brain Sciences*. Vol. 20(4), pp. 537-96.
- Raubenheimer, D., Bernays, E.A. (1993) "Patterns of feeding in the polyphagous grasshopper *Taeniopoda eques*: a field study", *Animal Behavior*. Vol. 45, pp. 153-67.
- Real, L.A. (1991) "Animal choice behavior and the evolution of cognitive architecture", *Science*.

- Vol. 253, pp. 980-6.
- Rolls, E.T. (2000) *The Brain and Emotion*. Oxford: Oxford University Press.
- Rolls, E.T. (2000) "Precis of 'The Brain and Emotion'", *Behavioral and Brain Sciences*. Vol. 23, pp. 177-233.
- Rosch, E. (1973) "Natural Categories", *Cognitive Psychology*. Vol. 4, pp. 328-50.
- Stotz, K. and Griffiths, P. (2001) "Dancing in the Dark: Evolutionary Psychology and the Argument from Design", in *Evolutionary Psychology: Alternative Approaches*. Scher, S and M. Rauscher (eds), Dordrecht: Kluwer.
- Thelen, E. (1995) "Time-scale dynamics and the development of an embodied cognition", in *Mind as Motion: Explorations in the Dynamics of Cognition*. R.F. Port, T. van Gelder (eds), Cambridge, MA: MIT Press.
- van Gelder, T. (1995) "What Might Cognition Be, If Not Computation?", *The Journal of Philosophy*. Vol. 92(7) pp. 345-381.
- van Gelder, T. (1998) "The Dynamical Hypothesis in Cognitive science", *Behavioral and Brain Sciences*. Vol. 21(5) pp. 615-627.
- Xu, F., and Carey, S. (1996) "Infants' metaphysics: the case of numerical identity", *Cognitive Psychology*. Vol. 30 pp. 111-53.