Wayne Christensen
Macquarie University
wayne.christensen@gmail.com

The decoupled representation theory of the evolution of cognition – a critical assessment

Abstract

Sterelny's Thought in a Hostile World (2003) presents a complex, systematically structured theory of the evolution of cognition centered on a concept of decoupled representation. Taking Godfrey-Smith's (1996) analysis of the evolution of behavioral flexibility as a framework, the theory describes increasingly complex grades of representation beginning with simple detection and culminating with decoupled representation, said to be belief-like, and it characterizes selection forces that drive evolutionary transformations in these forms of representation. Sterelny's ultimate explanatory target is the evolution of human agency. This paper develops a detailed analysis of the main cognitive aspects. It is argued that some of the major claims are not correct: decoupled representation as defined doesn't capture belief-like representation, and, properly understood, decoupled representation turns out to be ubiquitous amongst multicellular animals. However some of the key ideas are right, or along the right lines, and suggestions are made for modifying and expanding the conceptual framework.

1 Introduction

There is a bewildering range of ideas on the evolution of cognition. An intuitive and widely held schema for the evolution of cognition is that environmental complexity selects for behavioral flexibility, which in turn requires some kind of key representational ability. There are, however, many different ways that this schema can be applied. Many theories attempt to cast the non-cognitive/ cognitive distinction in terms of a contrast between reactive and anticipative creatures. For instance, Dickinson and Balleine (2000) contrast habit machines with capacity for goal-directed action, which they find in rats. This idea is connected to the widespread view in animal cognition research that cognition is to be defined as, or at least can be diagnosed in terms of, behavior control that can't explained in terms of basic associative learning mechanisms (Zentall 1999). Another kind of distinction used to frame the evolution of cognition is between domain specific and domain general abilities; a major transition is thought to occur from creatures that only rely on domain-specific abilities to those capable of cross-domain integration. Menzel and Giurfa (2001), Spelke (2003) and Premack (2007) all appeal to this contrast, but whereas Menzel and Giurfa see flexible integration in honeybees, Spelke and

Premack think that domain general cognition is uniquely human. Spelke proposes that the first mechanism supporting domain-general cognition is language.

Ideas about specific representational mechanisms are also varied. The most general definition of cognition with positive content sees it as some kind of information processing that intervenes between receipt of the stimulus and the response (Shettleworth 1998; Zentall 1999). Tomasello and Call (1997, p. 11) regard the basis for cognitive flexibility as lying with internal models that can manipulated in advance of interaction with the world. Language is a perennial favorite explanation for human uniqueness, but other proposals include hierarchical 'mental construction' abilities (Gibson 2002) and capacity for analogical reasoning (Gentner 2003). Herrmann et al. (2007) even suggest that there are no significant cognitive differences between humans and apes (other than social), and that culture accounts for enhanced human cognition.

Thus, not only is there quite a range of different specific ideas on the evolution of cognition, there are diverse ways of conceptualizing it. In this context Sterelny's account is notable because it directly and systematically addresses many of the core issues, offering an account of the nature of the selection that drives cognitive complexification, and a taxonomy of complexification culminating in belief-like representation. The account also analyses a transition from drive-based motivation to motivation involving preferences, and in combination these elements provide the core elements for understanding the origins of belief-desire psychology, central to many conceptions of human agency.

Sterelny's account is therefore an important touchstone for theorizing the evolution of cognition, and what follows is a detailed analysis of the main cognitive aspects.[1] The core structure of the theory is outlined in section 2, and section 3 examines in more depth the taxonomy of behavior control forms and the analysis of the selection that drives transformation in behavior control. Sterelny's account is built around the idea that belief-like representation is associated with a late evolutionary separation of indicative and imperative function; this is questioned, and reasons are given for thinking that complexification of perception-behavior relations should occur much earlier than Sterelny's account suggests. Section 4 considers a range of evidence in support of this, including sensori-motor complexification in early neural evolution, and evidence for both an advanced kind of decoupled representation (termed 'model-based representation') and preferences in rats, drawing on transitive inference, maze and incentive revaluation

[1] Sterelny also provides an account of hominid evolution which will not be examined.

experiments. Section 5 integrates this evidence into a revised theoretical picture that clarifies the sources of selection pressure and forms of control involved in the evolution of cognition. With basic versions of the cognitive abilities underlying beliefs and desires present in rats, it is suggested that, rather than being associated with the appearance of new kinds of representational abilities, as proposed by Sterelny and many others, human agency may be an endpoint of a long progressive elaboration and refinement of cognitive abilities supporting flexible goal-directedness.

2 The core structure of the theory

In thumbnail sketch Sterelny's theory is as follows. The account begins with Godfrey-Smith's (1996) analysis of selection for flexible control of behavior, which says that flexible response to environmental variation is adaptive when the benefits of detecting the variation and responding to it outweigh the costs. Sterelny's theory describes a succession of increasingly complex forms of sensorimotor control, beginning with a *detection system* baseline involving specific adaptive response to a specific environmental signal (p. 14)[2]. He defines a *detection agent* as an organism equipped only with detection systems (p. 14). The next grade of complexity is *robust tracking*, which involves tracking important features of the environment using multiple cues (p. 17). The culminating grade, *decoupled representations*, are "...internal states that track aspects of our world, but which do not have the function of controlling particular behaviors." (p. 29). The selection pressure driving this transformation is *translucency*, defined as the condition where the functionally relevant features of the environment map in complex ways onto the cues the organism can detect (p. 21). These perceptuo-motor forms appear progressively, with decoupled representation being associated with social cognition in great apes. This empirical distribution lends plausibility to the claim that decoupled representation is the basis for beliefs in humans. In addition to these stages of cognitive complexification Sterelny gives an account of the evolution of motivation describing two stages of motivational complexification: drives and preferences. Drives are a non-representational form of motivation control, whereas preferences represent goals for action and can be formed and modified through learning (pp. 92-95). Sterelny argues that drive-based motivation will be inadequate when the animal's behavioral repertoire include many options and it must access a wide range of ecological resources (pp. 92-94). He claims that this transition has occurred in the human lineage but is incomplete.

---

[2] Unless otherwise noted all page numbers for Sterelny are to Thought in a Hostile World.

I now explicate the account in more detail, schematizing the main elements and formulating the account in terms of models in order to make the overall structure clearer and to provide points of reference for comparison.

2.1 The Environmental Complexity Thesis

Sterelny's account begins with Godfrey-Smith's (1996) account of the evolution of behavioral control, which Sterelny describes as a framework (p. 11). Godfrey-Smith conceptualizes the idea that cognition evolves in response to environmental complexity as the *environmental complexity thesis* (ECT), according to which the function of cognition is to enable the agent to deal with environmental complexity. To make this more precise he develops an analysis of the circumstances under which environmental complexity will select for flexible behavior (1996, ch. 7). This analysis poses the problem in terms of the two possibilities (1996, p. 207) depicted in figure 1. In the baseline state the animal's sensory discriminations don't distinguish between distinct environmental conditions, and the animal's behavior must be adaptive across these hidden variations. Effectively the animal employs a fixed behavioral strategy in varying conditions. In the alternative state the animal discriminates the variations in environmental state and produces behaviors matched to each condition. The question is what conditions will favor 1b over 1a. Sterelny summarizes Godfrey-Smith's account of these conditions as follows (2003, p. 13):[3]

1. The organism's environment varies in ways that matter to that organism.
2. The organism has relevant variation in its repertoire; different actions have different payoffs in different environments.
3. The organism has access to information about its environment.
4. The benefit of optimizing behavior to the specific state of the environment outweighs its costs, including the cost of error. There must be a major fitness difference between optimum and satisficing responses, or the agent must have a reliable signal of environmental differences.

[Figure 1 about here.]

2.2 The evolution of beliefs and desires

After outlining the Environmental Complexity Thesis Sterelny elaborates an account of the evolution of complex cognition, with the ultimate aim of understanding human agency (p. 4). He addresses what are standardly seen as the two fundamental components of human agency: beliefs and desires. In part his account examines the status of folk

[3] Sterelny's theory is the focus here so it is primarily this explication that matters for what follows.

psychology, but this aspect is largely bypassed here since it makes little difference to the core structure of the theory. Sterelny uses theoretical definitions of beliefs and desires, and places the burden of justification for these definitions on the overall success of the account (p. ix). Questions will be raised about these definitions below, but shifting the strategy of justification to the theoretical and empirical analysis of cognitive architecture is a reasonable move.

2.2.1 The evolution of decoupled representation

2.2.1.1 Formal structure of the account

Figure 2 depicts Sterelny's account of the stages of complexification leading to decoupled representation. This will be called representational/control complexification here because it straddles both representation and behavior control. Robust tracking increases the sensory pathways between a key environmental condition and behavior, whilst decoupled representation increases the number of behaviors to which a given sensory discrimination contributes.[4]

[Figure 2 about here.]

As Sterelny notes, on the face of it robust tracking and decoupled representation are independent kinds of complexification (e.g., p. 36). Thus, in principle decoupled representation could be based on detection, as depicted in figure 3a. In practice Sterelny thinks this situation is unlikely (pp. 31-32), and that the typical case will be as depicted in figure 3b, with decoupled representation based on robust tracking. This gives something of a stage-structure to the three forms, with robust tracking being a precondition for decoupled representation. Thus, the account can be interpreted as proposing the evolutionary model 2a → 2b → 3b.

Sterelny argues that the selection force driving this complexification is informational translucency. Translucency is defined in contrast with informational transparency: an environment is informationally transparent for an animal when there is a simple and reliable correspondence between sensory cues and the functional properties of the environment that determine success (p. 20). That is, 2a will be adaptive. It is translucent when the functionally relevant features of the environment map in complex ways onto the

[4] As a matter of terminology, Sterelny only explicitly accords the label 'representation' to decoupled representation, although elsewhere he gives a definition of representation that seems to fit robust tracking (2001, p. 211). To avoid prejudging key issues the relatively neutral term 'sensory discrimination' will be used as the basic descriptor here.

cues the organism can detect (p. 21). In this condition 2a is no longer adaptive, and selection will drive a population towards 2b and 2c/3b.

[Figure 3 about here.]

2.2.1.2 The phylogenetic distribution of traits

Sterelny gives many examples of detection-system behavior across a wide range of phyla, including bacteria (p. 14), plants (p. 14), arthropods, including cockroaches (p. 14), fireflies (p. 15) and spiders (p. 18), and vertebrates, including fish (p. 18) and birds (p. 18). In addition he characterizes Artificial Life/Artificial Intelligence situated agents as detection agents. He gives fewer examples of robust tracking, but the examples include arthropods (bee navigation, pp. 23-24), and vertebrates (Reed warbler detection of cuckoo eggs, p. 17, and Piping plover broken wing display, p. 27). This is consistent with the idea that robust tracking is more phylogenetically restricted than detection-system behavior.

With regard to evidence for decoupled representation, Sterelny considers navigation, tool use and ecological knowledge, and social cognition. A so-called ‘cognitive map’, or representation of the environment used for navigation, would be an example of a decoupled representation because the map can be used to guide more than one kind of behavior. In contrast, a so-called ‘route-based’ method of navigating doesn’t (at least on the face of it) use decoupled representations, because getting from point A to point B is represented in terms of specific behaviors. Sterelny finds no clear evidence for the use of cognitive maps in navigation, although he says that rats appear to show intelligent use of spatial information (p. 45). With regard to the use of tools, he claims that most animals appear not to have decoupled representations of tools (p. 48), meaning that they only discover that tools can produce particular outcomes by trial and error, rather than by means of insight into the causal properties of tools. He does however consider suggestive evidence that crows show some understanding of tools (p. 48). Similarly, he finds no unequivocal evidence for decoupled representation of ecological resource information. On the other hand he thinks it likely that great apes have decoupled representations of social information (pp. 52, 76), given the extensive cognitive demands of social life (pp. 53-76).

This assessment of the evidence is consistent with the 2a → 2b → 3b model, and with idea that decoupled representation is a late evolved, highly advanced capacity. The association with great apes puts it in the right territory to be the basis for the evolution of

human beliefs, because amongst primates apes have been considered to have human-like cognitive abilities distinctly more advanced than monkeys (e.g. Byrne 2000).

2.2.1.3 The explanation for this distribution

The key issue in applying the theory is determining the main factors that contribute to translucency. Sterelny points out that transparency can be achieved through adaptation, as organisms become tuned to their environment (p. 21). He identifies two main sources of translucency: ecological generalization, which involves exposure to a wider range of environmental conditions (pp. 23-4), and hostile interactions. He gives greater weight to hostility because other animals actively try to deceive and defeat (pp. 24-5), whereas the inanimate and non-hostile biological world is indifferent to the animal, and in some cases even cooperative (p. 25).

This analysis helps explain the pattern of evidence above if we assume that much of the animal world is specialized rather than generalized.[5] We would then expect most animals to use detection-based behavior control, and see robust tracking only among those whose adaptive strategy revolves around adaptability, or in other words, coping with environmental variability. But other animals will be tend to be more variable than the physical environment, and it's plausible that the complex competitive social worlds of the great apes are particularly difficult to 'read'. Since translucency will be high, these are circumstances where the theory says we should find decoupled representation.

2.2.1.4 The connection between decoupled representation and beliefs

No direct, explicit argument is given linking decoupled representation to beliefs, but in light of the conceptual and empirical structure of the account we can identify some bases for the association. Humans show substantially greater behavioral flexibility than other animals, and human beliefs can play a role in guiding indefinitely varied behavior. Conversely, in general, non-human animals are relatively inflexible, and we might suppose that an important part of the explanation for this is that their representations aren't able to guide behavior flexibly. Given these points, the suggestion that decoupled representation – representation that contributes to multiple behaviors – is 'belief-like' looks reasonable. The general idea that human cognition is in some way decoupled from the environment certainly matches many intuitions; for instance, Dreyfus (2007) and McDowell (2007) disagree on many points, but they agree that animals are captivated by

[5] In fact we need to assume more than this; a variety of complications are discussed in later sections.

their environments, whilst humans in contrast are free in the sense that they can step back and reflect.

2.2.2 The evolution of preferences

Figure 4 depicts schematically the two stages in Sterelny's account of the evolution of motivation. Drives compete with each other to control behavior, and being non-representational they control behavior 'directly' rather than through a cognitively mediated process. Each drive imposes a distinct pattern on sensory-behavior relations, amplifying some kinds of responsiveness and inhibiting others. In contrast, preferences influence behavior control based on representations of outcome value, and they are subject to cognitive revaluation (p. 92). Thus, with preference-based motivation novel things and behaviors can be assigned and reassigned motivational value based on cognitive assessment of circumstances.

[Figure 4 about here.]

With regard to the phylogenetic distribution of drive-based motivation, Sterelny says that "many organisms simply have a built-in motivational hierarchy"[6] (p. 81), and that "something like preferences have evolved in the hominid lineage, and perhaps others. But that transformation is very unlikely to be complete" (p. 95). Thus, the 4a → 4b transition is associated with advanced cognition.

He offers two main sources of support for this view. The first is a theoretical argument based on transparency: because animals are internally co-adapted we can expect that internal signaling will evolve towards transparency (pp. 80-81), and as a result drive-based motivation is likely to be a more reliable means of keeping behavior in line with value than are preferences (p. 85). Sterelny argues that drives are only inadequate if one or more of the following conditions are met: (i) the animal's behavioral repertoire includes many possible options, (ii) a wide range of resources is needed, which are hard to encompass with a repertoire of built-in drives, (iii) summation mechanisms for determining which drive controls behavior don't produce good results, and (iv) the sensory profile of resources doesn't stay stable over evolutionary time (pp. 92-94). He evidently believes that these conditions don't hold for most animals. It might have been expected that Sterelny would formulate these conditions in relation to translucency, as suggested in figure 4, but the connection is left unclear.

---

[6] Elsewhere he cites McFarland (1996) on this point (Sterelny 2001, pp. 248-9).

Buttressing the idea that conditions (i) - (iv) are rare is a skeptical analysis of Dickinson and Balleine's (2000) claim to have shown preferences in rats, discussed in detail in section 4.3 below. If Dickinson and Balleine were right that rats have preferences this would throw into doubt the empirical picture described in section 2.2.1.2 above. The apparent rarity of decoupled representation would seem to indicate that for most animal behavior directly cueing off perceptual stimuli is an effective form of behavior control. If this is the case there would seem to be no need for preferences. Decoupled representations are only needed when individual perceptual discriminations are insufficient to determine which behavior should be performed, and this would seem to be the same conditions in which preferences gain adaptive value. That is, flexible valuation is only needed when a given stimulus doesn't have a fixed implication for behavior control. So decoupled representation and preferences should arise in roughly the same circumstances, and indeed, we might think that decoupled representation would be the representational substrate for preferences. In rejecting Dickinson and Balleine's putative finding Sterelny thus removes what is potentially a significant anomaly.

As with decoupled representation, Sterelny doesn't provide a direct, explicit argument linking his conception of preference to human desires. The requirement that preferences be subject to flexible cognitive valuation doesn't appear to be true for many cases commonly thought of as desires; the desire for chocolate or heroin may be suppressed on given occasions, such that the options 'eat chocolate' or 'take heroin' are subject to cognitive valuation on particular occasions, but the dispositional desires for chocolate or heroin may be resistant to cognitive revaluation. This might be the kind of thing that Sterelny has in mind with the claim that the transition to preferences in humans is incomplete. Clearly, though, some human desires involve cognitive valuation, and the flexibility of human agency depends on this. Intuitively, humans take decisions: that is, decide on particular actions taking into account structured features of the situation. This essentially involves cognitive valuation, because the value of the options is assigned at the point of choice based on the structured relations. Focusing on cognitive valuation may thus be a useful approximation for getting at key features of human motivation.[7]

3 Clarifying the structure of the theory

To evaluate the account we need a clear understanding of three kinds of things: (i) the nature of the behavior control taxonomies, (ii) the taxonomic transitions that are possible

[7] As a side note, Sterelny's idea that human preferences are cognitively-based, and hence flexible, is reminiscent of the idea that human's are unique in having a capacity for second-order desires (Frankfurt 1971), but Sterelny doesn't address the issue of higher-order motivation per se.

or likely, and (iii) the factors that drive transitions. This section explores the structure of Sterelny's theory from an abstract perspective.

3.1 Initial taxonomic questions

In order to develop a better conceptual understanding of the forms of behavior control described above this section outlines several other forms of simple behavior control (figure 5) and compares them to the taxonomies of figures 1 and 2.

3.1.1 Integrative behavior control

A very simple way to make behavior flexible is to make behavior production sensitive to the recent history of responses, as with habituation and sensitization (figure 5a). In the case of habituation, production of the response declines with repeated presentation of the stimulus, whereas in the case of sensitization it increases. In 5b a second signal is involved in the control of the behavior; B occurs as a consequence of the conjunction of the two signals. In 5c the second signal serves as a contextual modifier. 5c is similar to 5a inasmuch as the effect is modulation of a first order relation, but it differs in that the modulatory signal comes from beyond the immediate signal-behavior mechanism.

[Figure 5 about here.]

An initial question to ask is how these forms of behavior control relate to those of figure 1. In particular, can they be thought of as variants of 1b? 5a and 5c seem to fit, because in each case there is the discrimination of an additional condition which is used to provide control for behavior modification. 5b doesn't quite fit because the additional discrimination doesn't provide control for a behavioral modification, it refines the production of a given behavior.

These comparisons help to make it clearer that complexification can take some subtle forms. We might simplistically think that increase of behavioral flexibility will occur through the addition of a second complete signal-behavior pathway, but 5a and 5c highlight the possibility that behavioral flexibility can be achieved through the addition of regulation to a given pathway. 5b makes it apparent that controlling behavioral flexibility isn't the only adaptive reason for making additional discriminations. In the case of 5b the additional discrimination serves to make production of the behavior more focused.

The last point has implications for Godfrey-Smith's analysis of the evolution of behavioral complexification. This analysis is intended to specify the conditions in which

it is adaptive for the organism to discriminate environmental complexity, or as Godfrey-Smith puts it, answer the question "[w]hen should environmental complexity bring it about that the organic system will make a distinction, will attend to a difference in the world?" (1996, p. 11). Thus, Godfrey-Smith's account is based on the model of perceptual complexification shown in figure 6a. However in light of 5b we can identify 6b as a distinct model of perceptual complexification. Godfrey-Smith's account could be used to predict 5a and 5c, but it doesn't predict 5b. If 5b is an empirically significant form of behavior control, as it surely is, we should prefer the 6b model of perceptual complexification. This model draws our attention to the problem of behavior targeting as a general issue that can take a number of forms. That a given perceptual signal can adequately target a given behavior is taken for granted by the conceptualization of the problem of perceptual complexification schematized in figure 1. However this isn't always the case; multiple perceptual signals may be needed for the effective targeting of an action.

[Figure 6 about here.]

Of course, using multiple signals to control behavior is just what Sterelny's concept of robust tracking is about. However the distinction between 6a and 6b raises questions about the relationship between Sterelny's account and that of Godfrey-Smith. Although Sterelny describes Godfrey-Smith's analysis of the evolution of behavioral flexibility as a framework, it isn't clear exactly how this is so. On the face of it Godfrey-Smith's stage 2 (1b) is different to Sterelny's stage 2 (2b). Moreover Sterelny's stage 2 looks like it's driven by a different problem — that of targeting a given behavior — than Godfrey-Smith's stage 2, which is driven by the problem of flexible control. A plausible interpretation is that Sterelny sees Godfrey-Smith's account as explaining how detection systems evolve, and his account as covering, not the pressures that add further detection systems, but selection for more complex forms than detection. An explicit analysis of the relations between the two accounts would be helpful, however.

Are the forms of behavior control shown in figure 5 cases of robust tracking? They each involve the use of multiple signals to control a behavior, and at one point Sterelny defines robust tracking as the ability to use several cues to control behavior (pp. 28-29). However Sterelny at another point says that robust tracking systems track "important features of [the] environment" (p. 17), and this is shown in 2b. But there is no requirement that the multiple signals of 5b and 5c have a single environmental source. Indeed, for both 5b and 5c S1 might have an external source and S2 might have an internal source, or vice versa. In the case of 5c contextual control signals will generally come from a different source to

the primary signal whose effect is modulated; this is part of their functional value. There are at least two different classes of things that can be tracked: (i) particular environmental (and internal) states and entities, and (ii) the conditions for behavior production. Much of the way that Sterelny talks about robust tracking suggests that he has (i) in mind, but, as discussed below, the account treats (i) and (ii) as effectively equivalent. Yet (i) and (ii) can come apart, and 5c in particular looks like it's aimed at regulating behavior production rather than tracking some particular thing in the world.

3.1.2 Behavior management

Figure 7 depicts three forms of behavior management. As evolutionary complexification proceeds, and animals gain an increasing variety of behaviors, the need for coordination between behaviors increases. Behavior management mechanisms reduce this problem by providing coordination and resolving conflict. Figure 7a shows an elementary reciprocal modulation arrangement in which activation of one pathway influences the other; these links can inhibitory or excitatory. Conceptually it parallels 5a, inasmuch as the modulatory signals are internal. In the case depicted in 7b, S3 is a context signal which favors B1 in some circumstances and B2 in other circumstances; this parallels 5c. 7c shows a more complex arrangement involving a specialized arbitration system. One advantage of a specialized behavior management system like 7c, as compared with 7a and 7b, is that the arbitration itself can be context-sensitive. Thus, whereas 5a, 5c, 7a, and 7b all show first order control flexibility, 7c shows second order control flexibility. Like 7b, 7c incorporates a context signal, but it also illustrates several more sophisticated behavior management mechanisms. Prospective arbitration receives sensory signals, anticipates response conflict, and inhibits one or more conflicting behaviors. Retrospective arbitration detects the actual occurrence of response conflict and arbitrates the conflict. These two mechanisms can be coupled via learning: retrospective arbitration can manage novel response conflicts, and learning can transfer retrospective to prospective control.

[Figure 7 about here.]

It isn't clear how behavior management relates to Sterelny's taxonomy of figure 2. As was noted above, he defines a detection agent as an organism equipped only with detection systems. However he also says that detection agents can learn (pp. 14, 17) and have motivation systems (p. 20). This doesn't seem to be entirely coherent. If an agent has a motivation system then at least some of its behaviors are being governed by multiple signals. Depending on the interpretation of robust tracking this might count as robust tracking, but it doesn't fit the definition of a detection system: a modulated

sensory-behavior relation (5c) is different to an unmodulated sensory-behavior relation (2a). It is also hard to see how detection agents could learn since all their perceptual discriminations are coupled to behaviors. Learning requires flexibility in the coupling between perception and behavior, and Pavlovian learning will yield robust tracking, not detection, as Sterelny himself notes elsewhere (2001, p. 271).

Sterelny discusses Artificial Intelligence/Artificial Life "situated agents" as examples of detection agents (pp. 18-19), citing Brooks (1991b). However, although it is true that the behavior-based design approach involved cueing behavior off the environment to the greatest extent possible, the architectures nevertheless incorporated carefully crafted internal systems for managing interactions between behaviors. In the 'subsumption architecture' a given behavior module can inhibit other modules, and relationships between modules are organized hierarchically to achieve increasingly complex high level behavior. Brooks (1991a) describes a somewhat more sophisticated behavior management system inspired by animal hormone systems and applied to a walking robot called Attila. The simulated hormone system provided a centralized communication system coordinating behaviors according to high-level states such as 'sleep' and 'fear'. As Brooks notes, the point of this system is that a given behavior *doesn't* respond to a perceptual stimulus in the same way every time.

Sterelny makes the same point in relation to motivation: "[A]n agent's actions depend partially on its internal environment. An animal that registers increasing dehydration will behave differently from one that is satiated. A warm animal will not act like one that is cold" (p. 79). Yet transparency has been defined as the condition in which responding to a cue results in successful behavior, and detection systems are correspondingly defined as being adaptive under conditions of transparency (pp. 20-21). If it is adaptive to respond differently to a given cue in different circumstances then these are not conditions of transparency, and the response mechanism is not a detection system.

This lack of clarity makes it hard to know what detection agents really are, and what their evolution might involve. The most parsimonious interpretation of Sterelny's position is that in addition to detection systems detection agents have drive-based motivation systems, this is their only kind of behavior management (compare 7b with 4a), and the only departure from transparency is variation internal state. This rules out learning, but allowing detection agents to learn requires a major reconceptualization of what they are. For instance, if an animal needs to learn then it isn't living in conditions of transparency.

A further point can be gleaned from the issue of behavior management. As was said earlier, the complexification taxonomy of figure 2 straddles representation and control. Coupled with the taxonomy of figure 4, Sterelny's overall picture of cognitive complexification can be called a *two stream view*: representational/control complexification and motivational complexification form two main streams of cognitive complexification (figure 8a). The behavior management mechanisms characterized in 7c aren't specifically motivational in nature, however, and research in artificial intelligence and neuroscience has suggested that behavior management mechanisms other than narrowly motivational ones play an important role in the evolution of cognition (Bryson 2001, Prescott 2007). This suggests a *three stream view* of cognitive complexification (figure 8b), with the evolution of specialized control mechanisms being partly independent of motivation and control.

[Figure 8 about here.]

In summary, the additional forms of behavior control canvassed here help us to pose questions about the taxonomies of figure 1 and figure 2. The behavior control forms of figure 5 highlight the significance of regulation and suggest that there is more to perceptual complexification than the flexible control of behavior, or in other words that we should reject Godfrey-Smith's formulation of the problem driving the evolution of cognition (6a) in favor of a more generalized formulation in terms of action targeting (6b). They also highlight ambiguities in the nature of robust tracking, and in particular the distinction between tracking particular states and entities, and tracking the circumstances for behavior production. The importance of regulation is reinforced by considering behavior management (figure 7). This finds no easy home in Sterelny's taxonomy and should perhaps be considered a different kind of complexification.

Sterelny doesn't say that he's identifying all the kinds of complexification there are, and identifying phenomena he doesn't talk about isn't an objection to his account per se (though it might count as a useful extension). As will be discussed below, however, control complexification poses specific conceptual problems: Sterelny can't endorse a three stream picture and be consistent.

3.2 Initial selection questions

Uncertainty about taxonomy will lead to uncertainty about which transitions can occur and what factors drive them. To get a better understanding of transition issues we need to turn to the analyses of the selection pressures.

3.2.1 Flexible control and internal complexity

With some qualifications Godfrey-Smith's conditions for the evolution of flexible control appear plausible. It is hard to see how flexible control can evolve if these conditions aren't satisfied, and if they are satisfied then it appears that flexible behavior control will be selectively favored over an inflexible 'catchall' behavior. One qualification has been raised above: as an account of the basis for cognitive complexification Godfrey-Smith's conceptualization is too narrow. The control of behavioral flexibility is just one of a broader class of behavior targeting problems that can drive cognitive complexification. A second qualification concerns the formulation of the problem in terms of environmental variation. There is nothing essential to the payoff structure for flexible control which requires that the variability be *environmental.* We could also talk about behavior-significant conditions *elsewhere in the agent,* or even just *elsewhen,* since the prior history of producing an action can affect whether and how it should be produced now. Even the current state of the behavior-producing mechanism *itself* can make a difference, in the sense that it influences what behavior is possible. If the muscles are depleted from recent exertion then climbing the steep hill may be out of the question; if the hand is grasping the fruit it can't be used to pick up the stick. We can also go a step further and question whether the analysis should be restricted to the flexible control of overt behavior, since internal processes can be flexibly controlled, and indeed, we might expect there to be a mutually supportive relationship between the flexible control of internal processes and the flexible control of overt behavior. For example, a primary function of the autonomic system is to flexibly adjust the body's physiology to the current needs of action (Powley 2003).

With a suitably liberal interpretation of 'environment' and 'behavior' Godfrey-Smith's analysis can encompass these phenomena, but in fact the account is framed in terms of variation of the external environment. Much of the first part of the 1996 book is devoted to discussion of the nature and relative merits of externalist, internalist, and constructivist kinds of theories[8], and in the context of these alternatives he presents an adaptationist externalism for mind as a bet (e.g., p. 56). This suggests that he doesn't think internal complexity will be be a significant part of the story. Thus, we can distinguish between two models of selection for flexible control. On Godfrey-Smith's model (figure 9a), which we can call the *environmental complexity model* (corresponding to his Environmental Complexity Thesis), the evolution of flexible behavioral control is driven by environmental variability. On the alternative model (9b), which we can term the

[8] Roughly speaking, the former focus on the role of the environment and the latter focus in various ways on internal structure and relations between the system and environment.

complexity model, the evolution of flexible control is driven by both environmental and internal variability.

[Figure 9 about here.]

What difference does the difference between these models make? Before the question can be answered directly a point of clarification is needed. In Godfrey-Smith's framework the environmental complexity model is also the main component of the Environmental Complexity Thesis, a proposal concerning general function of mind and the selection force that drives the evolution of cognition as a whole. But given the distinction between the two models of figure 6 we should be cautious about such generalizations. Selection for behavioral flexibility is an important part of, but not the whole of, the selection forces driving the evolution of cognition. Correspondingly, we shouldn't treat the complexity model of 9b as an alternative to the Environmental Complexity Thesis – a general theory of selection for cognition – but rather as a component of a larger account of the selection forces at work in the evolution of cognition. However 9b may nevertheless make significant difference to the way we understand the evolution of cognition, compared with 9a. Identifying internal complexity as a source of selection for flexible control could lead us to see selection for flexible control as stronger than otherwise, because we are including an additional source of complexity. We might also identify different kinds of flexible control than if we focus only on environmental complexity. An account that emphasizes internal complexity may see motivational complexification as more elaborate and occurring earlier, and it is likely to see a stronger role for behavior management, leaning more towards 8b than 8a.

3.2.2 Translucency

3.2.2.1 Complex mappings

Translucency is defined in contrast with informational transparency as follows (pp. 20-21):

> If the signal indicating the presence of a specific resource is reliable, and if the agent can use its sensory apparatus to discriminate that signal from other stimuli, then cue-driven behavior will succeed. For with respect to that feature, the animal lives in an *informationally transparent environment.* Detection systems reliably drive adaptive behavior in transparent environments. But cue-driven organisms will often struggle if ecologically relevant features of their environment — their functional world — map in complex, one to many ways onto the cues they can detect. Such organisms live in *informationally translucent environments*.

The first question to ask about translucency is how it relates to Godfrey-Smith's conditions for the evolution of behavioral flexibility. Since the next stage after detection is robust tracking, translucency appears to be a targeting problem for a particular behavior rather than a problem specific to flexible behavior control (figure 6). Translucency therefore can't be equivalent to the selection in Godfrey-Smith's account. Even so, there is some connection: condition 1 of the four conditions for selection for flexible control refers to environmental variability, and translucency refers to complex mappings from environmental conditions to the stimuli the animal detects. In both cases it seems that the environmental aspect of the problem involves complex mapping. Sterelny refers to *one-to-many* mappings from environmental conditions to signals detected, however, whereas the issue in Godfrey-Smith's account is a *many-to-one* relationship (figure 1a). It subsequently becomes apparent that Sterelny includes both kinds of mapping in translucency, since he says that "many different sensory registrations form a single functional category, and similar physical signals may derive from very different functional sources" (p. 21).

To clarify matters we can distinguish between *ambiguity,* where multiple environmental conditions map onto a given sensory discrimination, and *synonymy*, where a given environmental condition maps onto multiple sensory discriminations (figure 10). As far as sensory discrimination in general goes, both ambiguity and synonymy can come in good and bad forms. Ambiguity is good when there is no point in making more fine grained distinctions; this can also be called *optimal classification.* In part, Godfrey-Smith's account says that when classification is sub-optimally coarse grained (relative to the needs of flexible behavior control) selection will favor more fine grained classification. There is also the converse situation: when classification is excessively fine grained selection will favor more coarse grained discrimination. Since environmental conditions here are being defined as 'ecologically relevant' then ambiguous classification is presumably suboptimal: the animal would do better by making more distinctions. Thus, the problematic aspect of complex world-sensory mappings is sub-optimal classification. Sub-optimal for what purposes? As has just been noted, for Sterelny's account it is the targeting of particular actions that matters, so we need to set aside condition 2 of Godfrey-Smith's criteria for the evolution of flexible behavior. This makes it clearer that translucency is not the same as selection for flexible behavior.

[Figure 10 about here.]

Synonymy in general can also be good or bad. Detecting an environmental condition in more than one way can provide greater reliability, and in fact this an advantage of robust

tracking that Sterelny emphasizes. In the case of detection systems synonymy looks like a problem because the animal is responding to one and the same environmental condition with different behaviors. This isn't necessarily a problem, though, since the environmental condition might warrant more than one behavior. We have to assume that each animal-relevant environmental condition warrants one and only one behavior, distinct from the behaviors warranted by other animal-relevant environmental conditions, before synonymy is necessarily a problem. This won't always be the case. Bad (or nonfunctional) synonymy will occur when only some of the behaviors cued by the environmental condition are appropriate to it. This will be a kind of excessively fine grained classification: the animal makes more distinctions than it should.

Thus, translucency could be defined as suboptimal classification for the purposes of action targeting. In the case of ambiguity the classification is too coarse grained, whereas in the case of synonymy the classification is too fine grained.

3.2.2.2 How does translucency select?

It might seem unnecessary to ask how translucency selects, because if it is suboptimal classification then, obviously, forms of discrimination that provide improved classification will be selectively favored. It's important to understand how translucency will select for robust tracking and decoupled representation in particular, however. The form of translucency that Sterelny initially identifies, here termed non- or dysfunctional synonymy, doesn't look like it will select for either: synonymy should select for pruning. It might give rise to robust tracking because the wiring transformation is simple, as depicted in figure 11, but it doesn't select for robust tracking as such.

[Figure 11 about here.]

On the other hand, suboptimal ambiguity will select for robust tracking when integrating over a second signal will reduce the ambiguity, and figure 12 sketches a possible transformation from suboptimal ambiguity to robust tracking using this kind of mechanism. EC1 is an adaptively important condition for the animal, but the sensory discrimination S1 is ambiguous between EC1 and EC2. The addition of a second layer of discrimination integrates S1 with a second perceptual signal sourced from EC1. S4 detects the conjunction of S1 and S2, and as a result correlates better with EC1 than does S1.

[Figure 12 about here.]

This is only one of a variety of kinds of robust tracking that can reduce ambiguity, and the claim that suboptimal ambiguity can select for robust tracking seems safe. In general, failures of recognition will manifest as kinds of ambiguity: false positives are an ambiguity of the tokening of the perceptual state, whilst false negatives are an ambiguity of the non-tokening of the state. Introducing a second perceptual signal isn't the only way to reduce ambiguity because this can also be achieved by tuning the individual sensory signals. But given that there will be limits on the extent to which a given perceptual signal can be tuned, the reduction of ambiguity through multiple signals should be a favored form of organization in many circumstances.

What of decoupled representation? Why would ambiguity or synonymy give rise to representations that contribute to more than one behavior? Sterelny says that "many translucent environments select for decoupled representation, for in many such environments information becomes available in a piecemeal fashion and without its immediate significance for action being apparent" (p. 78). Thus, the idea seems to be that in certain situations a given item of perceptual information won't uniquely determine which action should be performed (figure 13a). Presumably, in such circumstances, in order decide on a particular behavior the animal must use additional information (13b). This tells us that an animal may need more than one item of information to decide on a particular behavior, but it doesn't tell us why a given item of information would play a role in more than one behavior. Sterelny doesn't spell this out, but he may have in mind the possibility that a given item of information, conjoined with different other signals, can point to different behaviors. Thus, in 13a S1 is ambiguous between B1 and B2. Coupled with S2, as in 13b, S1 points specifically to B1. But perhaps, when coupled with S3, S1 points to B2 (13c). Since S1 can point to both B1 and B2 (13d), it counts as a decoupled representation.

[Figure 13 about here.]

Framing the point more generally, decoupled representation can occur when perceptual discrimination has a finer granularity than whole behaviors. Combinations of discriminations are needed to decide on a particular behavior, different combinations point to different behaviors, and a given discrimination can participate in multiple combinations, thereby pointing to different behaviors in different circumstances.

3.2.2.3 Indication uncertainty versus control uncertainty

Perception can be thought of as uncertainty reduction, but comparison of figures 11 & 12 versus 13 suggests that somewhat different kinds of uncertainty are at work in these cases. The problem driving 11 & 12 is the isolation of a specific environmental condition, whereas the problem driving 13 is isolating a specific behavior. In section 3.1.1 it was said that robust tracking is ambiguous between tracking particular environmental states and entities, and tracking the conditions for behavior production. Translucency shows essentially the same ambiguity: it is defined as a complex mapping between ecologically relevant features of the environment and the cues the animal can detect (pp. 20-21), but since detection and robust tracking both control individual behaviors, the complex mapping is also between environmental conditions and behavior.

We can clarify this ambiguity by distinguishing between indication uncertainty and control uncertainty (figure 14). Indication uncertainty is uncertainty about the source of a perceptual signal, whilst control uncertainty is uncertainty about the behavior to be performed.

[Figure 14 about here.]

It isn't an accident that translucency is ambiguous between indication and control uncertainty. Sterelny follows Millikan in assuming that much animal perceptual discrimination is behavior-specific, and that consequently these discriminations combine indication and behavior control functions. Millikan says: "Simple animal signals are invariably both indicative and imperative" (1989, p. 296), and Sterelny paraphrases this as: "when organisms are only equipped with relatively simple means of representing their world, we cannot draw a distinction between representations that merely report how the world is, and representations that direct behavior" (p. 29). Thus, in these cases uncertainty about the way the world is and uncertainty about what to do are effectively equivalent: the animal is built such that the resolution of the first kind of uncertainty also resolves the second. Millikan goes on to say that "The step from these primitive representations to human beliefs is an enormous one, for it involves the separation of indicative from imperative functions of the representational system" (p. 29). This looks very much like the agenda for Sterelny's theory of decoupled representation: explain the evolution of human beliefs in terms of the separation of indicative and imperative functions. In Sterelny's taxonomy indicative and imperative functions only clearly come apart in the case of decoupled representation.

However conceptualizing translucency as an undifferentiated amalgam of indication and control uncertainty may gloss over important distinctions. In principle we could expect that the reduction of control and indication uncertainty might occur in quite different ways, and thus have different evolutionary effects. Although some formulations of robust tracking are ambiguous (e.g., p. 28), the gist of it seems aimed at indication uncertainty. Many mechanisms for regulation, on the other hand, are aimed at reducing control uncertainty, and the control forms of figure 5 can be interpreted this way. In the case of 5b, if S1 and S2 have their origins with different sources then it looks like the mechanism is squarely aimed at control uncertainty reduction. If they come from a common source then the mechanism is more ambiguous as between indication and control uncertainty reduction. Viewed as mechanisms for control uncertainty reduction, the problem in 5b and 5c is that S1 alone doesn't fully reduce control uncertainty. The second signal provides additional information that can reduce control uncertainty further. Looked at this way, it can be a distinct advantage if S2 comes from a different source in each case, because this can yield greater information and hence better restriction of behavior production. Greater restriction corresponds to more precise targeting.

We can codify these distinctions in terms of two models of the evolution of behavior targeting (figure 15). In the model of 15a, targeting problems are resolved through the reduction of indication uncertainty. In the model of 15b, targeting problems are resolved either through indication uncertainty reduction or control uncertainty reduction. The ungainly but accurate term multiconditionalization describes improved behavior targeting through the conditionalization of behavior on multiple states or entities (for instance, respond to the red light only in the white room). When S2 comes from a distinct source to S1, 5b and 5c count as forms of multiconditionalization (as does 5a, where the additional restriction is provided by a history signal).

[Figure 15 about here]

In Sterelny's overall account it appears that the model of 15a applies to most animal behavior. That is, in most cases behavior targeting is improved by reducing indication uncertainty. This can't be the case for decoupled representation, however, because decoupled representations don't uniquely point to specific behaviors. When decoupled representations are being employed some mechanism other than indication uncertainty reduction must be used to resolve which behavior to perform. Indeed, Sterelny's account of the evolution of decoupled representation (p. 78) suggests that it is driven by need for multiconditionalization; individual items of information don't determine which behavior

should be performed (13a & 13b). Thus, it appears that on Sterelny's account the model of 15b only comes into play with the emergence of decoupled representation.

Another way to describe this is that indication and control uncertainty come apart for decoupled representation. Sterelny's account implies the model shown in figure 16. Indication and control uncertainty are effectively equivalent for both detection systems and robust tracking, and are distinct in the case of decoupled representation. Phylogenetically, this separation occurs late and is restricted to advanced cognition.

[Figure 16 about here.]

In light of the model of figure 16, Sterelny is committed to the late, restricted emergence of multiconditionalization as a behavior targeting strategy, associated with decoupled representation. Strictly, although he doesn't mention the possibility, his account can allow for limited kinds of multiconditionalization prior to the appearance of decoupled representation, employing what could be termed pluri-coupling: multiple perceptual discriminations are coupled to a given behavior, none of which link to any other behavior. Multiconditionalization could only occur by means of behavior-specific perceptual discriminations because any perceptual discrimination that links to more than one behavior will count as a decoupled representation.

When posed in this way Sterelny's account is arguably implausible on abstract grounds. Multiconditionalization appears to have relatively simple implementation requirements, and is a powerful mechanism for behavior targeting. It should therefore evolve early and be widespread. Multiconditionalization based on pluri-coupling is consistent with Sterelny's account, but if pluri-coupling is pervasive we might expect selection to favor decoupling, since if a perceptual discrimination is in place it is efficient to re-use it in the control of other behaviors. Multiconditionalization should consequently give rise to decoupled representation, which will then also appear early. Of course, whether the implementation requirements of multiconditionalization really are simple depends, in part, on how easy it is to construct perceptual discriminations that link to more than one behavior. If this is difficult then this constraint might impose rarity on multiconditionalization.

4 Revising the empirical picture

The abstract structure of Sterelny's theory has been considered at some length, and it has been suggested that the late separation of indication and control uncertainty is

implausible. If his empirical survey is right however, then the models of figure 2 & figure 16 are supported by the evidence. This section re-examines the empirical picture.

4.1 The early evolution of decoupled representation

The early evolution of nervous systems is a useful point of comparison for Sterelny's account of representational complexification. As a very simple multicellular animal without a nervous system, sponges provide the baseline state. In sponges cells called myocites perform both sensory and motor functions (figure 17a), contracting to regulate water flow through the body of the sponge. Because they perform both sensory and motor functions myocites are called *independent effectors* (Swanson 2003, p. 16). The next grade of complexity is illustrated in Cnidaria, where sensory and effector functions have separated (figure 17b). Sensory neurons, derived from ectoderm, have specialized for sensory discrimination. This allows much greater specialization for sensory functioning, and it allows the cells to be positioned in locations best suited for sensing, as opposed to motor actions. But in addition the separation of function allows a given sensory neuron to innervate multiple effector cells (called *divergence*), and it allows a given effector cell to receive input from multiple sensory cells (called *convergence*). This is shown in figure 17c. These properties allowed dramatic advances in sensorimotor complexity, which included specialized sensory systems dedicated to particular classes of stimuli, such as chemicals, temperature and light (Swanson 2003, p. 21).

[Figure 17 about here.]

Comparison of figure 17c with 2b and 2c suggests that proto or elementary versions of both of Sterelny's grades of representational complexity are present in the Cnidarian nervous system. In 17c B1 takes input from S1 and S2, so has at least part of the structure of robust tracking, and S1 innervates B1 and B2, thus having the form of decoupled representation. Calling different effector cells different behaviors might seem a stretch, but there is no obvious reason to rule this out and it is a short step from divergent innervation of effector cells to influencing multiple distinct overt behaviors (illustrated in examples discussed below).

Convergence and divergence form the basis for further sensorimotor complexification, and both robust tracking and decoupled representation can be seen more clearly in more elaborate, yet still relatively basic, cases. Figure 18a shows an intuitive example of a core perceptual mechanism, *hierarchical feature analysis*, from Hubel and Wiesel's (1962) account of hierarchical feature construction in the cat visual cortex. Neurons in the lateral

geniculate nucleus (LGN) are sensitive to dot shaped stimuli. Neurons in the striate cortex (or V1) called *simple cells* take input from multiple LGN neurons and are sensitive to oriented lines. It is not depicted in 18a, but Hubel and Wiesel showed an additional stage: some simple cells integrate from LGN cells sensitive to dark dots, and hence are sensitive to dark oriented lines, whilst others integrate from cells sensitive to light dots, and are sensitive to light oriented lines. *Complex cells* integrate across light- and dark-sensitive simple cells, and are sensitive to lines of a particular orientation, whether light or dark. Downstream areas of the visual system extract more complex features: for instance in primates neurons in MT are sensitive to motion information of various kinds, whilst neurons in the temporal cortex are sensitive to object features and identity. Visual processing in primates is organized into major partially parallel streams, beginning with M, P and koniocellular pathways from the retina through LGN and extending to the dorsal and ventral streams (Reid 2003).

[Figure 18 about here.]

Hierarchical feature analysis is an efficient and powerful way to construct more elaborate forms of perception from the starting point seen in Cnidaria. Individual receptor cells are the basic sensing unit on which complex perception is based, which means that the discrimination of complex stimuli must be achieved through comparison across multiple receptor cells. A given receptor cell is often highly ambiguous with respect to possible external signal sources, but this ambiguity can be reduced by downstream integration, as shown in 18a (and 12b). By iterating integration across multiple stages, increasingly complex high order information can be extracted, corresponding to increasingly complex features of distal sources. The relevant point here is that such feature representations will generally count as cases of both robust tracking and decoupled representation. The robust tracking aspect is apparent from 18a: the features are differentiated by means of multiple pathways back to the source. The decoupled representation aspect isn't diagrammed, but there are strong reasons to think that many feature representations can contribute to multiple behaviors. In the case of vision, e.g., many states and entities will be represented in terms of combinations of location, movement, form, and color information. Likewise, in the case of gustation many items experienced orally will be represented in terms of combinations of sweet, salty, sour and bitter features. Each feature will participate in the representation of many different specific things, and thereby contribute to different behaviors. In other words, a given feature can point to different behaviors as part of different feature complexes, just as depicted in figure 13.

Hierarchical feature integration probably emerged early in the evolution of specialized perceptual systems, in part because of the constraints of perceiving complex stimuli with individual receptor cells as the base sensory unit. Comparison across receptors then provides the only way to discriminate features more complex than individual receptors can detect. In the case of eyes, the organization of the eye, retina, and downstream pathways enables highly structured comparisons. This integrative organization begins very simply; as Land and Nilsson point out, "Even the simple pit eye of a planarian flat worm…has some ability to compare intensities in different directions" (2002, p. 4). The evolution of eyes was already well advanced by the early Cambrian (Schoenemann 2006), and Land and Nilsson suggest that visually guided predation may have been the trigger for the Cambrian explosion (2002, p. 2). This will have required relatively complex integrative information processing mechanisms because a mobile predator must identify its prey and contend with effect of its own movement on perception. 'Nystagmus', the ability to hold a fixed point of focus during movement and to rapidly shift to a new point, allows stable perception during movement. It is found in contemporary crustaceans, and Schoenemann argues that the stalk eyes of the crustacean *Leanchoilia*, found in the early Cambrian, may have supported nystagmus and, combined with strong swimming ability, allowed a predatory lifestyle (2006, p. 311).

Neural convergence is at least a necessary condition for robust tracking, and the evidence concerning nystagmus in *Leanchoilia* indicates that this potential was being exploited in relatively sophisticated ways in some of the earliest multicellular animals. Divergence supports decoupled representation, and by extending the argument at the end of the last section it can be predicted that there should have been an early elaboration of divergence to form increasingly complex, clearly articulated decoupled representations. Specialized neurons provide a long range, point-to-point communication mechanism that can serve either convergence or divergence just as easily. The appearance of specialized perceptual systems was, in effect, an adaptive investment in specialized mechanisms for indication uncertainty reduction. Distal perception requires both convergence and divergence, since convergence is needed for complex feature discrimination, and the effective discrimination of distal things requires flexible integration of feature information. Animals also have a pressing need for control uncertainty reduction, and convergence applied to motor control provides multiconditionalization, a powerful form of behavior targeting. Since perceptual information will often be relevant to more than one behavior, mechanisms for distributing perceptual information to multiple behaviors should be favored.

Pavlovian and operant learning are mechanisms of this kind. The perceptual discriminations that provide the perceptual raw materials for associative learning count as decoupled representations because, prior to learning, they aren't coupled to particular behaviors, and with ongoing learning they can be flexibly uncoupled (extinction) and coupled to new behaviors. The phylogenetic distribution of associative learning is very wide – for instance, the sea snail *Aplysia californica* is a model organism for research on the molecular mechanisms of associative learning (e.g., Kandel and Hawkins 1992) – which points to early evolutionary origins.

Taken together, these lines of evidence cast doubt on the stage structure of the 2a → 2b → 3b model – robust tracking and decoupled representation appear together, rather than in sequence – and they indicate that the empirical interpretation of the theory described in section 2.2.1.2 is mistaken. As predicted by the argument at the end of section 3.2.2.3, multiconditionalization and decoupled representation appear near the beginning of the evolution of multicellular animals, rather than near the end. The abstract structure of Sterelny's account of the conditions that select for decoupled representation is correct: information arrives in a piecemeal fashion, and may be relevant to more than one behavior (p. 78). However these circumstances aren't unique to the social worlds of great apes. In the early evolution of neurally-based sensorimotor systems information arrives in a piecemeal fashion, in part because the base sensory unit is the individual receptor cell, and because the discrimination of distal stimuli typically depends on integrating many different signals. The features discriminated can be relevant to multiple behaviors, so are functionally decoupled.

4.2 Relationally structured beliefs

If decoupled representation appears in the early evolution of nervous systems this puts into focus the question of whether the definition of decoupled representation adequately captures belief-like representation. As described earlier, the justification for this claim lies with the structure of the overall account and its empirical support. The passage where Sterelny discusses this is as follows (p. ix):

> "Belief-like" (and "desire-like") are somewhat weaselish expressions, and I confess to using them advisedly, to dodge controversies of which I want no part. … While I have an account of the evolution of cognitive states and a cognitive architecture that approximates these roles, that approximation is rough. Suppose that my picture of the evolution of the mind turns out, miraculously, to be true in every detail. To settle (say) whether my "decoupled representations" are really beliefs, we would need both a well-developed account of those folk commitments about belief, and a theory of reference for folk psychological vocabulary telling us

> the extent to which folk psychology's vocabulary depends on the accuracy of folk psychology's picture of the mind. While I have views on these matters, they are not the focus of this book. If my evolutionary scenario pans out, my suggestion is to think of the relationship between decoupled representation and belief as analogous to that between contemporary evolutionary accounts of species and speciation, and folk biological species concepts.

The important point here is that there is no *independent* support for the claim that decoupled representations are belief-like, other than the overall architecture and the evidence on which it is based. The idea that very simple neural phenomena fit the definition for decoupled representation is disorienting, and there is a temptation to think that these can't really be decoupled representations because they aren't very belief-like. But this would get the cart before the horse. It's a hypothesis that decoupled representation is a good definition of belief-like representation, and if such simple phenomena satisfy the definition this is counter-evidence to the hypothesis. Rejecting these cases as instances of decoupled representation because they aren't belief-like would beg the question.

Yet Sterelny reviewed research on navigation, tool use, ecological knowledge and social cognition, and found no clear evidence for decoupled representation, although he thought it plausible that great ape social cognition would involve decoupled representation. It is strange if decoupled representation can be found in very simple circumstances, and not in the case of more complex behaviors where it might be expected to play a stronger role. Arguably, however, evidence for decoupled representation can be found in these kinds of cases, when interpreted properly.

For instance, Dusek and Eichenbaum (1997) developed a transitive inference task for rats using spices buried in sand to provide distinct odors. The rats could dig in the sand and would in some cases find cereal. Training involved presenting rats with pairs of odors and rewarding according to the scheme A > B, B > C, C > D, D > E, where A > B means that the rat should choose odor A over odor B to find the food. In the test trial the rats were presented with the pair BD, which they had not previously encountered. To choose correctly they should pick B. In the experiment 88% of normal rats performed the transitive inference correctly. This appears to show decoupled representation: the new behavior is the correct choice in the novel situation, the pairing of B and D. The representation is the relational scheme extracted from prior training. The most interesting result from the experiment was that rats with a disconnected hippocampus performed at chance on the transitive inference probe, even though they learned the odor pairings

presented during training just as quickly as the normal rats. To put it another way, for rats with an intact hippocampus the BD pair was an extension of the problems they had been encountering recently, whereas for rats without a hippocampus the BD pair was a new problem for which they could only choose at random.

Notably, rats with hippocampus damage are also impaired at solving water maze tasks compared with normal rats. In a typical Morris water maze task the rat is placed in a pool in which there is a platform below the surface, which the rat can't see because the water is opaque. The pool is placed in a room with landmarks, and the rat must swim around until it finds the platform. The rats are released from varying starting positions during training. Normal rats are able to solve the problem, evidently learning to locate the platform using the configuration of landmarks, but rats with a damaged hippocampus can only learn to locate the platform if they are released from a fixed starting point, apparently able to associate the platform location with a single landmark, but unable to flexibly locate the platform within a landmark array (Eichenbaum 2000).

Taken together, this evidence suggests that the hippocampus is involved in extracting stable relational information that allows flexible problem solving, whilst rats without a hippocampus can only learn fixed responses. The relational information is a kind of decoupled representation, and it contributes to flexible behavior control. Above it was argued that the perceptual discriminations that participate in associative learning count as decoupled representations because they aren't coupled prior to learning, and can be re-coupled with ongoing learning. The decoupled aspect is easy to overlook because, once learned, the effect of the associations on behavior control is inflexible.[9] That is, the association is something like an acquired detection system. The kind of learning demonstrated here shows a more powerful kind of decoupled representation which can flexibly influence dynamic behavior control. Moreover, the hippocampus is associated with declarative memory, and the results are consistent with Eichenbaum's theory that declarative memory is constructed by extracting relational information across multiple experiences (Eichenbaum 2001).

In psychological research declarative memory has been distinguished from implicit memory (Cohen and Squire 1980), and is associated with conscious control and reasoning. It is thus, plausibly, the cognitive basis for what we usually think of as beliefs.

---

[9] At least, it is so in the case of 'habit learning', which is the traditional interpretation of all associative instrumental learning. Modern accounts recognize that not all instrumental learning is inflexible habit learning (Bouton 2007, pp. 403-414). 'Incentive revaluation' experiments, discussed in section 4.3 below, are one of the main sources of evidence for the contemporary view.

If the evidence described is correct then Sterelny's intuition that there is some important form of representation supporting beliefs is right, and it is correct that this kind of representation has a distinctively flexible role in behavior control. However the definition of decoupled representation doesn't effectively pick out this kind of representation. Eichenbaum's theory of declarative memory is one alternative cognitive model of 'belief-like representation'.

The overall picture is thus not as strange as it might have seemed. It would have been very odd if decoupled representations appeared in simple perception and behavior control, but were absent in more advanced behavior. However decoupled representation can be found in more complex animal behavior, and in fact we can distinguish multiple kinds of decoupled representation. There is the basic form of decoupled representation that participates in simple associative learning, and a more sophisticated, relationally structured kind of decoupled representation involved in flexible behavior control. This will be termed 'model-based representation' here because the extraction of stable relationships is effectively model building.

4.3 Evidence for preferences in rats

Evidence for a relatively sophisticated form of decoupled representation in rats has implications for Sterelny's arguments concerning rat preferences, because decoupled representation and preferences should arise in roughly the same circumstances, as discussed in section 2.2.2. This link is apparent in the reasons Sterelny gives for doubting that rats have preferences, where he says that the problems faced by rats don't require "broad responses" (associated with decoupled representation). Sterelny says that the rat world may be complex in the sense that knowing when intervention is needed is a difficult problem, however "the interventions are few in number, simple, and structureless" (p. 92). It's not entirely clear how to interpret these claims, but it looks like the transitive inference experiments described above are counter-evidence. The rats have 'broad responses' in the sense that they can appropriately adjust their behavior to a new problem that is related to previous problems they have encountered. The behaviors are structured in the sense that they are sensitive to problem structure.

In section 2.2.2 it was said that humans make decisions – they choose particular actions taking into account structured features of the situation – and this involves cognitive valuation. In light of what we've just considered it can also be noted that decisions, in this sense, require model-based representation. Further, we can draw a distinction between flexible valuation based on habit learning (Dickinson 1985), that doesn't involve decision

because option values are incrementally modified based on their experienced outcomes, and cognitive valuation involving decisions based on model-based representation. If rats have model-based representation then they are likely to make decisions, in the sense specified, and they should have preferences, in the sense of cognitive representations of outcome value. It's important, then, to consider Sterelny's arguments against preferences in rats to see whether they tell against this conclusion.

At this point some clarification is needed. To review, on Sterelny's taxonomy there are two forms of motivation, drives (4a) and cognitively based preferences (4b), and he specifies four conditions which should select for a 4a → 4b transition: (i) the animal's behavioral repertoire includes many possible options, (ii) a wide range of resources is needed, which are hard to encompass with a repertoire of built-in drives, (iii) summation mechanisms for determining which drive controls behavior don't produce good results, and (iv) the sensory profile of resources doesn't stay stable over evolutionary time. Condition (iv) suggests that preference-based motivation will be advantageous in conditions where learning is advantageous, and Sterelny's notion of drive-based motivation appears to be something like a specific hungers picture which involves minimal learning (see e.g. p. 86; compare Bouton 2007, p. 330). One of the major drive theories, that of Hull, incorporates learning, however. On Hull's account motivational state multiplies with learning to produce behavior strength (Bouton 2007, p. 331), and motivational state acts as a reinforcer: behaviors that reduce need are strengthened (Bouton 2007, p. 251). For a Hullian drive system there is thus no need for the sensory profile of resources to be stable over evolutionary time. Conditions (i) and (ii) are also consistent with Hullian drive; condition (ii) because by means of secondary reinforcement an animal can learn to value resources not covered by its repertoire of primary reinforcers. Hull's theory doesn't involve preferences, so conditions (i), (ii) and (iv) don't distinctively select for preference-based motivation.

This leaves condition (iii); if we incorporate Hull then the condition will refer to summation over option values determined by drive state and prior reinforcement learning. This kind of summation won't be very effective if the problem is sensitive to structured relations, and across multiple choices the impact of the structure varies. A hypothetical maze problem provides a simple example: if a rat's goal is to reach the end of the east arm of a cross-shaped maze, and it always starts from the south arm, then the structure of the maze drops out of the problem because over repeated choices the same option is always correct (turn right). If the rat starts from varied locations within the maze then the maze structure doesn't drop out of the choice problem. Each time the rat enters the center of the cross it has three choices: left, right, or straight (setting aside the options of staying

still or going back). Which choice is correct depends on the specific starting point on the current run, so to get it right the rat must keep track of it's current relation to the maze structure.

To make this more explicit, for simplicity we can set aside drive state and consider summation based on simple associative learning (i.e., without model-based representation). Such a mechanism will assign value to options based in their individual success history (e.g., left: successful 3 times, unsuccessful 1 time; right: successful 2 times, unsuccessful 4 times; straight: successful 1 time, unsuccessful 1 time), and sum over these values to arrive at a decision. This mechanism will tend to get it wrong because the summing doesn't take into account the fact that the success of particular options is sensitive to structured relations that vary across trials (e.g. the choice 'left' starting from the north arm produces a different outcome to 'left' starting from the east arm). Thus, the rat has a decision problem, on the definition given earlier, because at the point of choice it must take into account structured relations between multiple factors.

How does this relate to preferences? Sterelny says that preferences are representational and subject to cognitive revaluation (p. 92). As argued at the end of section 2.2.2, a decision problem requires the agent to value choice options based on representations of the problem structure, which means in effect that preferences are needed for decision problems. In learning theory the motivational value conferred on a behavior by the outcome it produces is known as *acquired motivation* (Bouton 2007, p. 342), so preferences in this sense could also be termed 'cognitively acquired motivation'.

If decision problems correspond to the breakdown of summation as an effective decision mechanism, then Sterelny appears to be right that condition (iii) selects for preferences (with the appropriate modifications to encompass learning). The question, then, is whether condition (iii) applies to rats. The hypothetical cross-maze example suggests that it does, because it's a variation on the variable-start water maze problem described above, at which rats are demonstrably successful. The example is also closely related to the 'place/response' task of Packard & McGaugh (1996). Here, rats are trained from a constant start position in a cross-maze to go to a particular arm baited with food. In the original experiment the rats received four trials of training a day over 14 days. On the 8th and 16th days probe trials were given in which the rat was placed in the opposite arm. A rat that goes to the same arm from the new location was said to show a 'place' strategy, whereas a rat that makes the same turn from the new location (and ends up at the wrong arm) was said to show a 'response' strategy. Normal rats showed a place strategy on first probe trial, and a response strategy on the second probe trial. In contrast, rats with an

impaired hippocampus showed no strategy on the first probe trial, and a response strategy on the second probe trial.[10]

The constant-start constant-target cross maze problem doesn't require decision because there is a fixed relationship between choice options and outcomes. Understandably, the rats don't know that this is the problem they have, and in the early stages of learning normal rats use a strategy appropriate to decision problems. In the later stages of learning a simpler strategy takes over. This is evidence indicates that rats do have preferences, understood as cognitively mediated valuations of choice options.

The main empirical basis of Sterelny's skeptical view of rat preferences is a critique of experiments reviewed in Dickinson and Balleine (2000). In these experiments rats are trained to press a lever for a novel food and then given an injection of lithium chloride, which induces gastric illness. If the rats are placed in the skinner box without being re-exposed to the food and allowed to press the lever (no food is delivered), they continue to press as before. On the other hand, if the rats are re-exposed to the food before testing, then placed in the box and allowed to press the lever, lever pressing is reduced (p. 195). In other words, rats have to experience the food again before the initial association between the food and the illness can have an effect on behavior (the aversion). Importantly, the rats don't actually receive the food during the test, so the reduced lever pressing isn't the result of directly experiencing the now devalued food as a consequence of lever pressing. The standard interpretation is that rats are anticipating the food outcome and transferring the reduced motivational value of the food to a reduced motivational value for the behavior that produces the food (Bouton 2007, pp. 404-409).

To show that it is the learned incentive value, and not the experience of nausea that generates the food aversion when it is manifested, Dickinson and Balleine administered a nausea-reducing drug at the time of test, the point after re-exposure when the animals normally manifest the aversion. If the aversion is produced by the experience of nausea when exposed to the food, the aversion should be reduced by the drug. However, as predicted by Dickinson and Balleine, the drug had no effect on the aversion. Sterelny objects that it is odd to refer to the acquisition of a food aversion as a change in preference, since in humans food aversions are cognitively opaque (2003, p. 84). He further objects that the failure of the drug to eliminate the aversion doesn't show the

[10] It isn't important to the current discussion, but the experiment also showed that rats with an impaired caudate nucleus showed a place strategy on both the first and second probe trials. Thus, a double dissociation between place and response strategies was found, with the hippocampus being necessary for the former and the caudate nucleus necessary for the latter.

aversion is based on a cognitive preference, rather, it might operate simply by making the food taste bad (p. 85), which presumably would be unaffected by the drug.

He is right that this specific experimental manipulation doesn't rule out this possibility, but the criticism doesn't address the main experiment, which shows that the rats alter behavior to obtain food based on changes in the value of the food (without direct experience of the changed behavior-outcome relation). The key issue in this experiment is the change in motivational value conferred on the *behavior* by the change in the motivational value of the food. Even if the motivational change in immediate response to the food is mediated by change in sensory response, the alteration in the motivational value assigned to the behavior must be cognitive. The experiments reviewed by Dickinson and Balleine (2000) fall within a larger class of *incentive revaluation* experiments that manipulate motivational outcomes and monitor the effect this has on behavior (Bouton 2007, pp. 404-409). Although aversion is a common method for producing outcome revaluation, other techniques are also used, such as satiation, and the appropriate alteration of behavior is robust across these different ways of producing revaluation.

Thus, Sterelny hasn't made a convincing case against Dickinson and Balleine's experiments, and there is a wider body of evidence that provides strong support for the idea that rats are capable of cognitively assigning motivational value to choice options.

5 A revised theoretical picture

Sterelny's highest grade of representational complexity is achieved by the early Cambrian, which means that the subsequent half billion years of cognitive evolution in animals – i.e., almost the entire course of multicellular animal evolution – has involved more complex phenomena than the theory can account for. The cognitive taxonomy must be modified and enriched, but the selection pressures too may be stronger and more complex than envisaged. It was suggested in section 3 that Godfrey-Smith's flexible control account of selection for cognition should be modified in favor of a generalized formulation in terms of behavior targeting, including both flexible control and the targeting of individual behaviors. The concept of translucency in effect makes this shift, since it is a problem applying to particular behaviors, but it could be clarified to more explicitly encompass behavior targeting in general. Such a reformulation might specify both an umbrella conception of behavior targeting problems, and particular targeting problems that favor particular kinds of representation and control. An important element of this would be recognition of multiconditionalization as a behavior targeting strategy.

Failing to recognize multiconditionalization is one of the most significant weaknesses in Sterelny's account; as argued above, we should expect that multiconditionalization evolves early and plays a profound role in cognitive evolution.[11]

The analysis of the sources of translucency also needs to be amended. Sterelny only identifies ecological generalization and hostility as factors contributing to translucency, and following the social intelligence hypothesis he places great emphasis on the complexity of the social worlds of the great apes. Perception of the inanimate and non-hostile biological world is argued to tend towards transparency. This makes almost no allowance for 'physical' problems of perception and behavior. Given the material and functional constraints of building perception from neural systems, many perceptual problems are difficult because of the physical problems of extracting the relevant information from a complex array of ambient signals. Similarly, there are many difficult 'physical' challenges in motor control and problem solving, such as solving problems with multiple related factors. These kinds of problems are surely major spurs and brakes on cognitive evolution; even humans are easily overwhelmed by problems with more than a handful of features, whether they are social or not.[12] In addition to the environment, the complexity of the organism itself makes a significant contribution to the problems of behavior control. It makes behavior targeting more difficult because for a more complex organism there are more options to be controlled, and more variation that affects perception and behavior success. The role of self-generated movement in making perception more difficult is a simple example. Importantly, this will select for enhanced perceptual and control abilities, as illustrated by the evolution of nystagmus.

With regard to architecture, Sterelny's account was characterized as a two-stream view of cognitive complexification (representation/control + motivation) in section 3.1.2, and contrasted with a three-stream view (representation + control + motivation). It was flagged there that control complexification isn't merely an issue that Sterelny doesn't address; he can't coherently recognize it as an independent stream. The reason for this becomes clear when we note that the postulated late appearance of decoupled

---

[11] Further evidence for the importance of multiconditionalization comes from the range of powerful effects that context sensitivity has on associative learning. One example is known as 'the context shift effect': a response learned in one context will generally be reduced in another context (Thomas 1985). Conversely, in the case of 'renewal' a response extinguished in one context is more likely to reappear in a different context (Bouton and Ricker 2004). 'Occasion setting' is a third example: a cue marks a context in which an instrumental behavior can be performed (Bouton 2007, p. 160).

[12] As illustrated by Raven's Progressive Matrices problems, often used in intelligence tests (Raven et al. 2003). With Miller's 'magic number' as a seminal example, cognitive psychology throughout its history has been concerned with information processing problems and constraints that are not specifically social in nature. This may have wrongly neglected social problems, but neglecting the body of evidence compiled by classical cognitive psychology, and assuming that non-social problems are easy, is unwarranted.

representation is also a late separation of indicative and imperative function. If the reduction of control uncertainty and indication uncertainty are generally welded together there is no room for independent control complexification, because control is automatically furnished by perception. On the other hand, if indication and control uncertainty separate early, as the evidence indicates, then there is a functional need for specialized control because an integration step is needed to bring perceptual information to bear on behavior. Increasing behavioral complexity will also favor control specialization, as behavior management mechanisms such as depicted in figure 7c are needed to handle response conflict and other behavior coordination problems. In addition, as relations between perception and control become more complex increasingly sophisticated valuation mechanisms will be needed to bridge the gap.

A relatively simple taxonomy covering the various lines of evidence discussed to this point identifies three kinds of control (figure 19). Fixed pattern control generates a stereotyped pattern that is not sensitive to learning. 'Cache-based' learning flexibly modifies the pattern of response over time, but is inflexible in immediate control: options are assigned a 'cached' value that is not recalculated at the time of choice. Model-based control is flexible both over time and in immediate control, where the value of options and the pattern of behavior is decided 'on the fly' based on models that capture aspects of the problem structure.[13]

[Figure 19 about here.]

In principle a given architecture might incorporate one, two, or all three forms of control (figure 20). When more than one form is present interactions between them can occur. For instance, learning can modulate fixed pattern controllers, modifying output. To avoid definitional confusion here we must note that what is fixed for a fixed pattern controller is the patterning generated by the controller itself. External modulation of the controller may result in altered patterning of the controller *output*, and if the fixed pattern controller is only one of several contributing to a particular behavior then the *behavior* may show labile aspects despite the fixity of the contribution from the fixed pattern controller. A controller itself can also have both fixed and flexible aspects. Cache-based and model-based controllers may cooperate or compete with each other and with fixed pattern controllers in the control of a particular behavior (see e.g. Daw et al. 2005, Yin and Knowlton 2006).

---

[13] The 'cache-based' and 'model-based' terminology follows Daw et al. (2005) and Sutton and Barto (1998). Cache learning is more commonly known as habit learning (Dickinson 1985).

[Figure 20 about here.]

A ‘three system’ control architecture (figure 20c) has an especially powerful and flexible way of managing control problems. Over evolutionary time fixed pattern control is adjusted to regularities stable across these timescales. In ontogenetic time cache-based control adjusts to capture regularities stable over extended periods of time. Some control problems, however, have variability in surface behavior-outcome relations but deeper structural regularities, and model-based control provides a mechanism for solving these kinds of problems.

With regard to model-based representation, the example discussed in section 4 was relationally structured memory dependent on the hippocampus, but there is likely to be more than one kind. Model-based representation is advantageous anytime one or more immediate perceptual signals are ambiguous and this ambiguity can be resolved using stored structured information. Locally ambiguous and incomplete information is endemic to perception; for instance objects are only seen from one perspective at a time, and are often partially obscured. Geon theory – the idea that objects are represented in terms of a basic set of geometric shapes (Biederman 1987) – illustrates the kind of role that model-based representation can play in perception, but more generally all sorts of frequently encountered perceptual relations might be captured in models. Model-based representation also plays a role in motor control, where forward and inverse models supplement the incomplete control information provided by perception (Wolpert et al. 1995).

Decision problems (as defined above) provide a starting point for understanding the kind of selection that would favor model-based representation and control. A decision problem can be thought of as a kind of translucency in which the mapping between correct choice and available options is variable across choices, yet still exhibits regularity. This is effectively what Shiffrin and Schneider (1977) referred to as a ‘variable mapping problem’ and used to study executive control processes in humans.

Together these additional factors provide a richer framework for understanding cognitive complexification. This is a starting point for understanding both complexification in early animal evolution, and why rats should turn out to have relatively sophisticated kinds of representation, control, and motivation. A complementary issue is why cognitive evolution should be so protracted. It is being suggested here that rats have versions of the cognitive abilities central to human agency, yet the agency of rats is manifestly nothing like as sophisticated as the agency of humans. The basic answer is probably just that the

problems are demanding, and the elaboration of increasingly refined abilities is an extended process. There is a tendency to understand the evolution of cognition in terms of the simple presence or absence of abilities, but this is clearly not the right picture for other kinds of traits. Around 380 million years of evolution separates the earliest terrestrial tetrapods – not unlike salamanders – and the cheetah, and it is clear that there are many challenges that must be overcome to get from a basic capacity for terrestrial locomotion on four limbs to the running abilities of a cheetah. A long progressive refinement of abilities may be as important in cognitive evolution as it has been in tetrapod locomotion. Concomitantly, trying to explain the distinctiveness of human agency in terms of the *appearance* of the basic agentic capacities involved might be a bit like trying to explain the running abilities of cheetahs in terms of the evolutionary appearance of legs.

More specifically, it is not implausible that selection for flexible goal-directedness plays a central role in the complexification that bridges rat-grade cognition and human-grade cognition. This is because flexible goal-directedness is an extended form of behavior targeting, and improved behavior targeting brings better adaptive returns. We might expect, then, that flexible goal-directedness will be a primary target of selection in cognitive evolution. But the cognitive abilities involved in goal-directedness are formidable, including the integration of multiple sources of information for decision, guided attention in a complex environment, mental models of the situation including task-relevant factors, selective memory retrieval, the ability to manipulate information in working memory, the ability to organize behavior to achieve particular outcomes, reorganize behavior according to changing task priorities, and so on. Rats do show basic forms of these capacities in mazes, including prospective memory for places not yet visited (see e.g. Zentall 1999), but it is not hard to see that scaling up might be a very extended process. Apes have much larger brains that allow generally greater information processing capacity (Roth and Dicke 2005), and, in particular, much more powerful executive control abilities than rats. This executive control capacity may be essential to harness their extended, flexible representational capacities. The explanation for their relative cognitive distinction may rest, then, not with a new kind of representational ability, but rather with a more integrated and powerful suite of cognitive abilities. Human cognition would be an extension of this trend. If this is right then we need to understand the major features of the evolution of human agency in terms of the coordinated evolution of an architecture.

6 Conclusion

It's been argued here that a number of aspects of Sterelny's account are not correct, but some important core ideas are right. Translucency, robust tracking and decoupled representation are all valuable concepts, and the amendments and extensions suggested here build on this. These include a generalized conception of behavior targeting, a concept of model-based representation, and a three-stream picture of cognitive complexification. Ultimately, for this kind of theory to prove its worth it must provide a unifying framework for empirical comparative cognition research, and I conclude by briefly suggesting how the ideas canvassed here could help integrate several common, unintegrated conceptions of the evolution of cognition. In the introduction it was said that ideas on the evolution of cognition are diverse. This is true, but there are some core ideas that are widespread, including (i) cognition is information processing more complex than can be accounted for by simple associative learning, (ii) cognition involves mental models, and (iii) cognition involves a transition from domain-specific to domain-general abilities. The taxonomy of figure 19 provides a basis for putting these ideas into a common conceptual framework. A simple transition from one kind of control to another is unlikely to capture very well either the appearance of cognition per se, or the appearance of advanced, human-like cognition. An architectural approach characterizing the elaboration of particular forms of control and interactions between multiple kinds of control would provide a more nuanced picture. A basic overall theoretical understanding would include an account of the core properties of each kind of control, an account of the properties of multi-system architectures, and an account of the factors that drive various kinds of elaboration of multi-system architectures. If the distinctive capacities of human agency lie with a particular kind of integrated architecture, as suggested above, then such theory will be an important part of a fundamental understanding of human agency.

Bibliography

Biederman, I.: 1987, Recognition-by-Components: A Theory of Human Image Understanding, Psychological Review 94, 115-147.

Bouton, M., and S. Ricker: 1994, Renewal of Extinguished Responding in a Second Context, Animal Learning and Behavior 22, 317-324.

Bouton, M.E.: 2007, Learning and Behavior: A Contemporary Synthesis, Sinauer, Sunderland.

Brooks, R.A.: 1991a, Integrated Systems Based on Behaviors, ACM SIGART Bulletin 2 (4), 46-50.

———: 1991b, Intelligence without Representation, Artificial Intelligence 47, 139-159.

Bryson, J.: 2001, Intelligence by Design: Principles of Modularity and Coordination for Engineering Complex Adaptive Agents, PhD thesis. Department of EECS, MIT,

Byrne, R.W.: 2000, Evolution of Primate Cognition, Cognitive Science 24 (3), 543-570.

Christensen, W.D.: 2007, The Evolutionary Origins of Volition, in D. Ross, et al. (eds.), Distributed Cognition and the Will, MIT Press, Cambrige, MA, pp. 255-288.

Cohen, N.J., and L.R. Squire: 1980, Preserved Learning and Retention of Pattern Analyzing Skill in Amnesics: Dissociation of Knowing How and Knowing That, Science 210, 207-210.

Daw, N., Y. Niv, and P. Dayan: 2005, Uncertainty-Based Competition between Prefrontal and Dorsolateral Striatal Systems for Behavioral Control, Nature Neuroscience 8 (12), 1074-10711.

Dickinson, A.: 1985, Actions and Habits: The Development of Behavioural Autonomy, Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 308 (1135), 67-78.

Dickinson, A., and B.W. Balleine: 2000, Causal Cognition and Goal-Directed Action, in C. Heyes, et al. (eds.), The Evolution of Cognition, Bradford Books/MIT Press, Cambridge, MA, pp. 185-204.

Dreyfus, H.: 2007, The Return of the Myth of the Mental, Inquiry 50 (4), 352-365.

Dusek, J.A., and H. Eichenbaum: 1997, The Hippocampus and Memory for Orderly Stimulus Relations, Proc. Natl. Acad. Sci. USA 94, 7109–7114.

Eichenbaum, H.: 2000, A Cortical–Hippocampal System for Declarative Memory, Nature Reviews Neuroscience 1, 41-50.

———: 2001, The Hippocampus and Declarative Memory: Cognitive Mechanisms and Neural Codes, Behavioural Brain Research 127, 199-207.

Frankfurt, H.G.: 1971, Freedom of the Will and the Concept of a Person, The Journal of Philosophy 68 (1), 5-20.

Gentner, D.: 2003, Why We're So Smart, in D. Gentner, et al. (eds.), Language in Mind: Advances in the Study of Language and Thought, Bradford Books/MIT Press, Cambridge, MA, pp. 195-235.

Gibson, K.R.: 2002, Evolution of Human Intelligence: The Roles of Brain Size and Mental Construction, Brain, Behavior, and Evolution 59, 10-20.

Godfrey-Smith, P.: 1996, Complexity and the Function of Mind in Nature, Cambridge University Press, Cambridge.

Herrmann, E., J. Call, M.V. Hernàndez-Lloreda, B. Hare, and M. Tomasello: 2007, Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis, Science 317 (5843), 1360-1366.

Hubel, D.H., and T.N. Wiesel: 1962, Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex, Journal of Physiology 160, 106-154.

Kandel, E., and R. Hawkins: 1992, The Biological Basis of Learning and Individuality, Scientific American 267 (3), 78-86.

Land, M.F., and D.-E. Nilsson: 2002, Animal Eyes, Oxford University Press, Oxford, UK.

McDowell, J.: 2007, What Myth?, Inquiry 50 (4), 338-351.

McFarland, D.: 1996, Animals as Cost-Based Robots, in M.A. Boden (ed.), The Philosophy of Artificial Life, Oxford University Press, Oxford, pp.

Menzel, R., and M. Giurfa: 2001, Cognitive Architecture of a Minibrain: The Honeybee, Trends in Cognitive Sciences 5 (2), 62-71.

Millikan, R.: 1989, Biosemantics, The Journal of Philosophy 86 (6), 281-297.

Packard, M.G., and J.L. McGaugh: 1996, Inactivation of Hippocampus or Caudate Nucleus with Lidocaine Differentially Affects Expression of Place and Response Learning, Neurobiology of Learning and Memory 65 (1), 65-72.

Powley, T.L.: 2003, Central Control of Autonomic Functions: Organization of the Autonomic Nervous System, in L.R. Squire, et al. (eds.), Fundamental Neuroscience, Elsevier Academic Press, San Diego, pp. 913-932.

Premack, D.: 2007, Human and Animal Cognition: Continuity and Discontinuity, Proc Natl Acad Sci USA 104 (35), 13861-13867.

Prescott, T.: 2007, Forced Moves or Good Tricks in Design Space? Landmarks in the Evolution of Neural Mechanisms for Action Selection, Adaptive Behavior 15 (1), 9-31.

Raven, J., C. Raven, and J.H. Court: 2003, Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1: General Overview, Harcourt Assessment, San Antonio, TX.

Reid, R.C.: 2003, Vision, in L.R. Squire, et al. (eds.), Fundamental Neuroscience, Elsevier Academic Press, San Diego, pp. 727-750.

Roth, G., and U. Dicke: 2005, Evolution of the Brain and Intelligence, Trends in Cognitive Sciences 9 (5), 250-257.

Schoenemann, B.: 2006, Cambrian View, Palaeoworld 15, 307-314.

Shettleworth, S.J.: 1998, Cognition, Evolution, and Behavior, Oxford University Press, Oxford.

Shiffrin, R.M., and W. Schneider: 1977, Controlled and Automatic Human Information Processing. Ii. Percepual Learning, Automatic Attending and a General Theory, Psychol. Rev. 84, 127-190.

Spelke, E.: 2003, What Makes Us Smart? Core Knowledge and Natural Language, in D. Gentner, et al. (eds.), Language in Mind: Advances in the Study of Language and Thought, Bradford Books/MIT Press, Cambridge, MA, pp.

Sterelny, K.: 2001, The Evolution of Agency and Other Essays, Cambridge University Press, Cambridge.

———: 2003, Thought in a Hostile World, Blackwell, Oxford.

Sutton, R., and A. Barto: 1998, Reinforcement Learning: An Introduction, MIT Press, Bradford Books.

Swanson, L.W.: 2003, Brain Architecture - Understanding the Basic Plan, Oxford University Press.

Thomas, D.R.: 1985, Contextual Stimulus Control of Operant Responding in Pigeons, in P.D. Balsam, et al. (eds.), Context and Learning, Erlbaum, Hillsdale, NJ, pp. 295-321.

Tomasello, M., and J. Call: 1997, Primate Cognition, Oxford University Press, New York.

Ungerleider, L.G., and M. Mishkin: 1982, Two Cortical Visual Systems, in D.J. Ingle, et al. (eds.), Analysis of Visual Behavior, MIT Press, Cambridge, MA, pp. 549-586.

Wolpert, D.M., Z. Ghahramani, and M.I. Jordan: 1995, An Internal Model for Sensorimotor Integration, Science 269 (5232), 1880-1882.

Zentall, T.R.: 1999, Animal Cognition: The Bridge between Animal Learning and Human Cognition, Psychological Science 10 (3), 206-208.

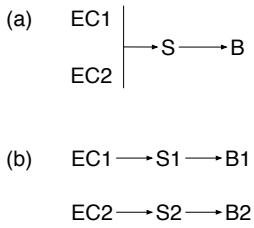

Figure 1: Godfrey-Smith's two stages of behavioral complexification. (a) Stage 1: indifference to environmental variation. (b) Stage 2: flexible response to environmental variation. B: behavior; EC: environmental condition; S: sensory discrimination.

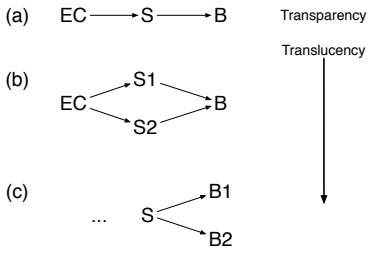

Figure 2: Sterelny's taxonomy of cognitive complexification. (a) Detection system. (b) Robust tracking. (c) Decoupled representation. Symbols as defined previously.

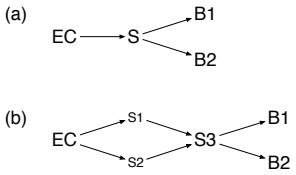

Figure 3: Two kinds of decoupled representation. (a) Decoupled representation based on detection. (b) Decoupled representation based on robust tracking.

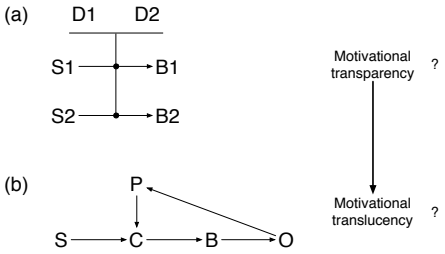

Figure 4: Sterelny's two stages of motivational complexification: (a) drives, (b) preferences. C: controller; D: drive; O: outcome; P: preference. Other symbols as defined previously.

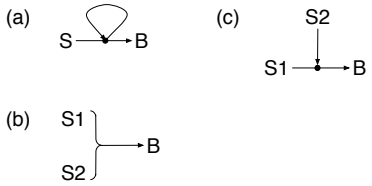

Figure 5: Three forms of integrative behavior control. (a) The propensity for a behavior is influenced by the history of behavior production. (b) B is activated by the conjunction of S1 and S2. (c) Context-modulation: S2 alters the S1—B relationship.

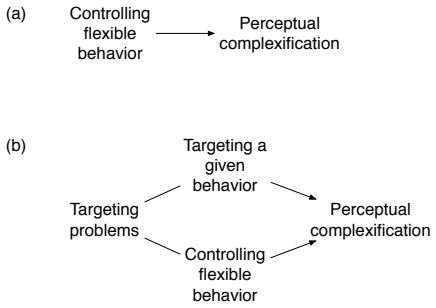

Figure 6: Two models of perceptual complexification. (a) Selection for perceptual complexification stems from the demands of the control of flexible behavior. (b) Selection for perceptual complexification stems from behavior targeting problems, which include both targeting individual behaviors and controlling flexible behavior.

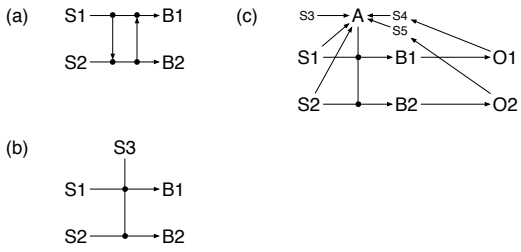

Figure 7: Three kinds of behavior management. (a) Reciprocal modulation; the links may be amplifying or inhibitory. (b) Contextual control. (c) Specialized arbitration incorporating context, prospective & retrospective arbitration. A: arbitration mechanism.

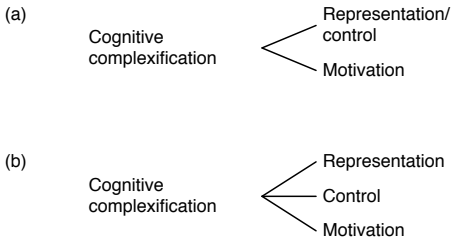

Figure 8: Two models of the main factors in cognitive complexification: (a) the two stream view, (b) the three stream view.

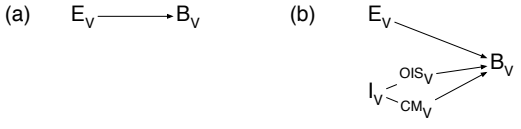

Figure 9: Two models of selection for flexible behavior control. (a) The environmental complexity model: variability in the environment drives the evolution of flexible behavior. (b) The complexity model: environmental and internal variability select for flexible control. B_V: variable behavior. CM_V: variation in control mechanism state; E_V: environmental variation; I_V: internal state variation; OIS_V: other internal state variation.

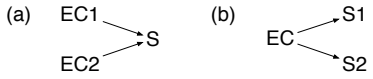

Figure 10: (a) Ambiguity, and (b) synonymy.

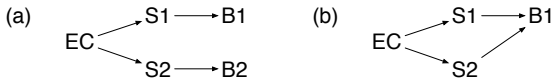

Figure 11: From non-functional synonymy (a) to robust tracking (b).

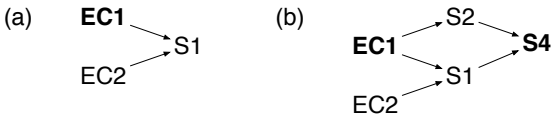

Figure 12: From sub-optimal ambiguity (a) to robust tracking (b).

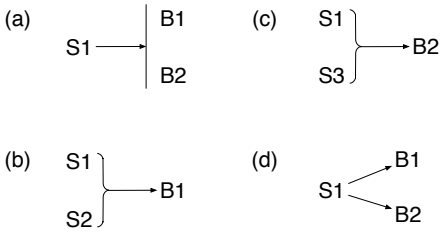

Figure 13: How behavioral ambiguity can give rise to decoupled representation. (a) Behavioral ambiguity. (b) A conjunction of signals resolves behavioral ambiguity. (c) A different conjunction of signals points to a different behavior. (d) The perceptual signal S1 can point to different behaviors, and hence is a decoupled representation.

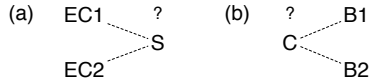

Figure 14. (a) Indication uncertainty: S is ambiguous between EC1 and EC2. (b) Control uncertainty: a controller C is uncertain between B1 and B2.

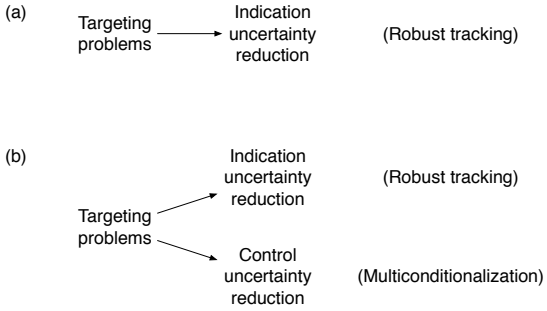

Figure 15: Two models of the evolution of behavior targeting. (a) Improved behavior targeting occurs through the reduction of indication uncertainty. (b) Improved behavior targeting can occur through the reduction of indication uncertainty or the reduction of control uncertainty

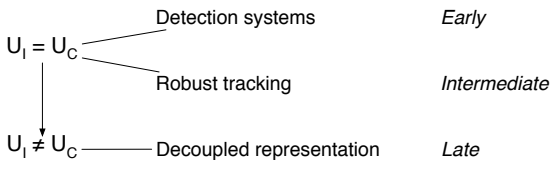

Figure 16: Sterelny's model of the separation of indication and control uncertainty. U_I: indication uncertainty; U_C: control uncertainty.

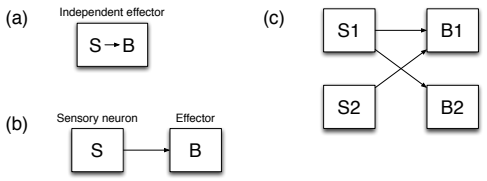

Figure 17: Separation of sensory and motor function in early neural evolution. (a) In sponges myocites perform both sensory and motor functions. (b) Cnidaria illustrate the next grade of complexity, with the separation of sensory and motor function. (c) Simple schematic depiction of convergence and divergence. Based on Swanson (2003), figures 2.5 and 2.6.

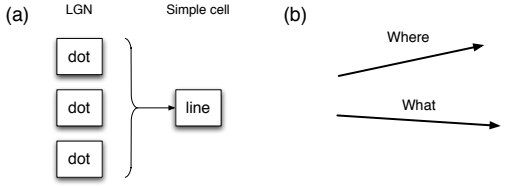

Figure 18: Hierarchical feature analysis in visual perception. (a) Discrimination of line stimuli in simple cells is achieved by integrating over LGN neurons sensitive to dot stimuli. (b) Feature extraction is elaborated into major streams; dorsal and ventral streams in macaques were originally characterized by Ungerleider and Mishkin (1982) as "what" and "where" pathways.

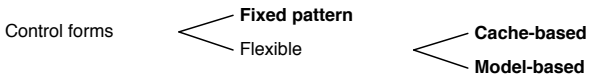

Figure 19: Three kinds of control.

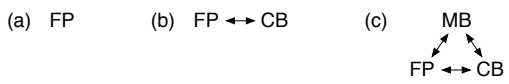

Figure 20: Three kinds of control architecture: (a) 'one system' fixed pattern architecture, (b) 'two system' architecture incorporating cache-based control, (c) 'three system' architecture incorporating model-based control. When there are multiple control forms they often interact in behavior control. FP: fixed pattern control; CB: cache-based control; MB: model-based control.