South African Journal of Philosophy

ISSN: 0258-0136 (Print) 2073-4867 (Online) Journal homepage: http://www.tandfonline.com/loi/rsph20

The Natural History of Desire

David Spurrett

To cite this article: David Spurrett (2015) The Natural History of Desire, South African Journal of Philosophy, 34:3, 304-313, DOI: 10.1080/02580136.2015.1057358

To link to this article: http://dx.doi.org/10.1080/02580136.2015.1057358

Published online: 15 Sep 2015.

South African Journal of Philosophy 2015, 34(3): 304–313
Printed in South Africa — All rights reserved

SOUTH AFRICAN JOURNAL
OF PHILOSOPHY
ISSN 0258-0136 EISSN 2073-4867
http://dx.doi.org/10.1080/02580136.2015.1057358

The Natural History of Desire[1]

David Spurrett
Philosophy; University of KwaZulu-Natal; Durban; South Africa
spurrett@ukzn.ac.za

Sterelny (2003) develops an idealised natural history of folk-psychological kinds. He argues that belief-like states are natural elaborations of simpler control systems, called detection systems, which map directly from environmental cue to response. Belief-like states exhibit robust tracking (sensitivity to multiple environmental states), and response breadth (occasioning a wider range of behaviours). The development of robust tracking and response-breadth depend partly on properties of the informational environment. In a *transparent* environment the functional relevance of states of the world is directly detectable. Outside transparent environments, selection can favour decoupled representations.

Sterelny maintains that these arguments do not generalise to desire. Unlike the external environment, the internal processes of an organism, he argues, are selected for transparency. Parts of a single organism gain nothing from deceiving one another, but gain significantly from accurate signalling of their states and needs. Key conditions favouring the development of belief-like states are therefore absent in the case of desires. Here I argue that Sterelny's reasons for saying that his treatment of belief does not generalise to motivation (desires, or preferences) are insufficient. There are limits to the transparency that internal environments can achieve. Even if there were not, tracking the motivational salience of external states suggests possible gains for systematic tracking of outcome values in any system in which selection has driven the production of belief-like states.

Introduction

In his *Thought in a Hostile World* (2003), Sterelny develops a detailed articulation of an approach suggested in Godfrey-Smith (1996, 2002). Godfrey-Smith's proposal, the Environmental Complexity Thesis (ECT), maintains that 'the function of cognition is to enable the agent to deal with environmental complexity'. The ECT proposes, that is, that organisms capable of cognition can respond more effectively to heterogeneity in their environments. For present purposes I simply accept the ECT. I happen to support it as a fruitful research programme, but will not offer a defence. My aim, rather, is to develop a line of thinking *internal* to the ECT, and which concerns the specific treatment of motivational states in Sterelny's (2003) book.

Sterelny describes his objective as combining two integrative projects that arise when humans are studied from a naturalistic perspective. One 'internal' project concerns the 'wiring and connection facts' about human cognitive architecture, and aims to assemble a 'coherent theory of human agency and human evolutionary history'. The other 'external' project attempts to relate the conclusions of the first project to the ways some social sciences (including psychology, anthropology, and economics) have produced the 'interpretation facts' which are 'refined versions of our folk self-conception', where that self-conception is that we are *intentional* beings (Sterelny 2003: pp. 3–5).

Sterelny frames his own proposal against the backdrop of a position he calls the 'Simple Co-ordination Thesis' (SCT). According to adherents of the SCT, which comes in various forms:

1 An earlier version of this paper was presented at the Annual Congress of the Philosophical Society of Southern Africa in Port Elizabeth, South Africa, 12–14 January 2015.

> …(a) Our interpretative concepts constitute something like a theory of human cognitive organization: they are a putative description of the wiring-and-connection facts; (b) Our interpretative skills depend on this theory, and our ability to deploy it on particular occasions; (c) We are often able to successfully explain or anticipate behaviour because this theory is largely true (Sterelny 2003: p. 6).

In the first part of *Thought in a Hostile World*, entitled 'Assembling Intentionality', Sterelny argues, in effect, that the Simple Coordination Thesis is *approximately* correct. He rejects the eliminativist view that belief and desire talk is false, and also the Dennettian attributionist view that the interpretation facts 'do not have the function of describing the internal organization of agents' (Dennett 1987; Sterelny 2003: p. 7). There are, Sterelny argues, internal 'belief-like' states that have features approximately like those expected by the SCT. They are elaborations of simpler systems, and it is unclear to what extent they are found in animals other than humans, but some other primates plausibly have them. In the case of preferences, Sterelny maintains that the SCT is less approximately correct, and 'desire-like' states incompletely found even in humans.

My critical concern is specifically with Sterelny's treatment of desire. Sterelny argues that there are important functional disanalogies between belief-like states and desire-like states, so that the considerations that explain why selection could in some cases favour the kinds of cognitive elaboration culminating in belief-like states largely *do not* apply to motivation. In his view, simpler control systems can achieve much more there, and so there is even less evidence that desire-like states are found in non-human animals. I begin with Sterelny's account of belief.

Sterelny on the descent of belief

Sterelny develops an evolutionary history of beliefs starting with the 'detection system,' an idealised and very simple control system that falls well short of belief. A detection system mediates 'a *specific adaptive response* to some feature (or features) of [an organism's] environment by registering *a specific environmental signal*' (Sterelny 2003: p. 14, emphasis in original). One of Sterelny's examples is the cockroach flight response, which triggers running away from gusts of air, registered by hair cells on their heads (p. 14). The idea is that the creature has a specific behavioural response (running away in this case) to a single environmental cue (the moving air caused by a striking toad, or magazine-wielding human). It seems uncontroversial that it could sometimes be a satisfactory solution to a control problem to have a behaviour triggered by a single cue.

It is not clear whether any specific organism *instantiates* a detection system strictly understood. At least some of Sterelny's examples are of animals whose flexibility in either detection or response is greater than that of a detection system as described. Sterelny also suggests that detection systems can be acquired by simple associative learning. For present purposes these worries can be set aside. The notion of a detection system is a useful *idealisation* even if there are no confirmed pure examples (Godfrey-Smith 2014: Chapter 2). Sterelny puts the notion of a detection system to work by thinking about the costs and benefits of such simple mechanisms, and possible forms of incremental modification that might lead to more discriminating control.

The most obvious benefit of detection systems is that they are relatively simple, and so cheap to build and run. As Sterelny points out, though, organisms with cue-driven behaviours can be vulnerable to exploitation. Fireflies which approach species-typical flash sequences to locate mates are lured by predators generating the same sequences to attract meals (Sterelny 2003: p. 15). Ants using the *absence* of chemical signals to distinguish (and attack) invaders are exploited by parasitic beetles mimicking signals and food-eliciting gestures (p. 15).

Sterelny refers to an environment in which signals that an organism can detect are reliably good occasions for specific responses it is capable of producing, that is where cue-driven behaviour will be successful, as *informationally transparent* (Sterelny 2003: p. 20). A transparent environment is 'characterized by simple and reliable correspondences between sensory cues and functional properties'. A key insight that he develops in his book is that not all environments have this property. In cases where relevant features of the environment 'map in complex, one to many ways onto the cues [an organism] can detect' then it occupies an *informationally translucent* environment

(Sterelny 2003: p. 21). In some cases the translucency is not the result of brute heterogeneity in the surrounding world, but is produced by other living things with an interest in misleading (for example by being camouflaged to fool predators or prey) in which case Sterelny calls the environment *informationally hostile*.

No general prediction follows from either transparency or hostility. But we can, Sterelny argues, make the *conditional* prediction that where the gains from more discriminating control outweigh the cost, then selection will favour certain kinds of elaboration of detection systems, if means are available. He discusses two particular kinds of elaboration—robust tracking, and response breadth.

Robust tracking is elaboration on the 'input' side. Where a detection system triggers a behaviour in response to a single cue, robust tracking links response to multiple, integrated cues. This can allow tracking of some environmental states under conditions of translucency or hostility. Reed-warblers are exploited by cuckoos, and face a serious problem distinguishing parasitic eggs from their own. Sterelny suggests that the multi-modal discrimination they draw on determining whether to reject an egg, including sensitivity to size, colour, shape, timing of appearance, and whether a cuckoo has recently been sighted near the nest, is an example of robust tracking (Sterelny 2003: pp. 27–29).

Response breadth, on the other hand, is elaboration on the 'output' side, and occurs when more than one behaviour might be produced in response to the same registered contingency. One of Sterelny's illustrations concerns responses to predators. Having registered the presence of a predator, an organism with response breadth might make one of a variety of responses including immediate flight, approach, or continuing with heightened vigilance, perhaps depending on the state of the organism itself (Sterelny 2003: pp. 33–40).

When robust tracking and response breadth are combined, we get what Sterelny calls 'decoupled representation'. When this happens, behaviour can be partly contingent on relatively high-level patterns of environmental information, perhaps integrated over time, and sensitive to the states of the behaving organism. Decoupled representations are 'internal states that track aspects of our world, but which do not have the function of controlling particular behaviors' (Sterelny 2003: p. 39). Such sophisticated states are, as found in humans at least, worth calling 'belief-like states'. These are genuine cognitive states which, while they may not share all of the features associated with any particular version of the Simple Co-ordination Thesis, are close enough that Sterelny's position regarding the interpretation facts in humans is neither eliminativist nor Dennettian attributionist.

Sterelny on the Descent of Preference

Sterelny devotes less attention to desire, or preference, than to belief. Three chapters of his book focus mostly on the natural history of belief-like states, followed by a single chapter on the descent of preference. Although there is some overlap, the comparative brevity of his treatment of motivation is not because it is a continuous elaboration of his account of belief. In fact, a major order of business in the motivation chapter is to argue that the account previously developed for belief *cannot* be generalised to motivation. The conclusion Sterelny eventually draws is, furthermore, more friendly to a kind of eliminativism. Although he finds a plausible rationale for the evolutionary development of belief-like states, he says that he does not 'think that there is even a rough mapping between preferences identified in our interpretive frameworks, and states of the internal cognitive architecture that controls human action' (Sterelny 2003: p. 87).

This is the conclusion that I wish to reject. I do so here by undermining Sterelny's argument that the belief treatment does not generalise to motivation. I therefore need to lay out Sterelny's reasoning in more detail than in the belief case. The criticism I offer here is both restricted and negative. I will not develop other lines of criticism of Sterelny's account of motivation, and will only be able to hint at an alternative positive view.

Sterelny's explanatory target

In the case of beliefs, Sterelny's explanatory target is relatively close to a standard (teleosemantic) conception. Belief-like states, as described, have representational content. They have satisfaction conditions and can be more or less supported by, and responsive to, environmental information. It is less clear that we are on such familiar ground regarding desires. Sterelny mostly does not

refer to 'desire-like states', and favours the term 'preference' for the motivational component of folk-psychological explanation. He describes his explanatory target as 'motivation based on representations of the external world' (p. 79). He seems, furthermore, to endorse the distinction drawn by Tony Dickinson between 'habit based' and 'intentional' agents, where intentional agents are sensitive to the value of actions, including values not explicitly cued by occurrent external information, *and* to the causal connections between acts and their consequences (e.g. Dickinson & Balleine 2000; Sterelny 2003: p. 82).

Sterelny thus associates 'preferences' with the capacity for means-end reasoning, suggesting that the 'most incontrovertible cases' of the applicability of belief and preference psychology are 'in complex calculating games like bridge and chess' (Sterelny 2003: p. 95), and that preferences as he understands them are to be distinguished from merely tracking the (possibly changing) values of world-states (Sterelny 2003: p. 86). Such cognitive and deliberative motivational systems are also, says Sterelny, to be distinguished from motivation by drives, or feeling. Drives, in his view, signal departures from homeostasis and in at least some cases motivate directly by feeling (although he does not say that *all* drive-based control involves feeling). He also maintains that drives can solve a wide range of control problems, and consequently that it is less clear that there is a job for preferences to do. As he poses the problem: '…what selective payoff could there be through routing action (say) through preferences about drinks rather than through sensations of thirst?' (Sterelny 2003: p. 81).

Various commentators have expressed dissatisfaction with how Sterelny opposes desire and preference-based motivation here (e.g. Schulz 2013; Papineau 2004). I will return to some of the difficulties with it in due course. For now, we need to be clear that Sterelny maintains that the motivational counterpart to beliefs is motivation based on representations of goal states and involving—or at least enabling—means-end reasoning. In his view for the evolution of preferences to be predicted, preferences need to do better (given the costs of implementing them) than motivation by drives signalling departures from homeostasis.

Why the belief case will not generalise

It might seem as though the considerations favouring the development of belief-like states would also explain the construction of motivational systems. Detection systems are 'pushmi-pullyu' (Millikan 1995) control solutions that yoke an *indicative* aspect (the environmental information to which each is sensitive) to an *imperative* one (the activity that each triggers). Decoupled representations replace these simple mappings with the more discriminating responses to environmental information that Sterelny calls robust tracking, *and* replace unique imperative output with response breadth—behaviour drawn from a set of possibly relevant activities. The latter elaboration, specifically, might seem by itself to create work for motivational states, to prioritise among the repertoire of available activities. This is the very inference that Sterelny wishes to block. The conditional argument from translucency or opacity to belief-like states *cannot*, he says, be generalised to give an account of 'motivation based on representations of the external world'. (Sterelny 2003: p. 79.)

The main reason Sterelny offers for this is that the departures from transparency that explain the existence of belief-like states are absent from internal environments. This is not, furthermore, accidental: Since internal environments have homogenous evolutionary interests, in the sense that all parts of an organism are—so to speak—on the same team, they both lack hostility, and will be under selection pressure for transparency. This means that signals of biological needs will tend to be trustworthy.

The natural physiological side-effects of departures from homeostasis have the potential to be recruited as signals for response mechanisms. Over time, we would expect these signals to be modified to become cleaner and less noisy; and internal monitoring systems to become more efficient in picking them up and using them to drive appropriate responses (Sterelny 2003: p. 80).

Drives might, of course, be simultaneously triggered in incompatible ways, but Sterelny maintains that one fairly robust solution to the control problem this poses is to have a 'built-in motivational hierarchy' (p. 81). He does not flesh this proposal out very much, but the idea seems to be that a relatively fixed ranking of drives can determine behavioural priorities in ways that depend

neither on representations of the values of outcomes, nor the connections between actions and their consequences. He refers approvingly to Rodney Brooks (1991) here, encouraging the suggestion that this fixed motivational hierarchy might depend to a significant extent on low bandwidth trumping relationships between drives. (In the early 1990s Brooks argued that 'intelligent' systems could be based on 'subsumption architectures' which were, roughly, hierarchies of detection systems operating without significant representational resources.)

Sterelny's positive view

While Sterelny holds that *many* organisms solve motivational problems without 'representing their needs', relying instead on a 'built-in motivational hierarchy' which ranks various drives mostly based on transparent internal signals, perhaps supplemented with external information, this is not true of *all* of them. The advantages of preferences over drives, according to Sterelny, include that they liberate motivation from 'immediate affective reward', that they allow more efficient decision making in cases where the range of available behaviours is large, that they allow a creature to have a smaller number of motivational states, that they permit motivational conflict resolution by means other than 'winner take all', and that preference-based systems are able to cope with changing needs, including needs that are phylogenetically novel (Sterelny 2003: pp. 92–95; see also Schulz 2013: p. 598). Preferences, as well as representing goal states, can be ranked, and they can be learned.

Sterelny's exposition is fairly cryptic on these points, and some commentators have expressed the view that the supposed advantages of preferences are not explained clearly enough, or that the contrast with what drive-based motivation could achieve is insufficiently motivated (again, see Papineau 2004). Certainly, Sterelny says very little to *justify* the claims that the number of motivational states would be smaller for drives than for preferences, or that drive-based motivation would *have* to resolve conflict on a winner-takes-all basis. Nonetheless he maintains that a small sub-set of species (hominids definitely, and maybe some others) are capable of more richly intentional and preference-based action, aimed at planning activities to achieve desired states of the external world. This, he thinks, does require some representation of value. But this transformation 'is very unlikely to be complete' (Sterelny 2003: p. 95). Sterelny maintains that much other behaviour allocation is likely based on fairly quick and dirty procedures and various kinds of distributed and non-representational control processes (as noted above, he explicitly and approvingly cites Brooks 1991).

The view that he reaches is, therefore, still a version of the environmental complexity thesis (ECT). The development of drive-based motivation, and of preferences, are alike in being responses to the problem of relating behaviour to external complexity. But, as Sterelny himself observes, his view of motivation is rather more friendly to eliminativism than his position in the case of belief, because the (alleged) transparency of internal environments means that there is rather less complexity for cognition to 'deal with'.

Criticisms of Sterelny

Sterelny, then, argues that internal environments will tend to be transparent, and that *because of this*, the inference from informational complexity to belief-like states does not generalise to the motivation. His argument distinguishes preferences from mere valuation, and focuses on preferences as requirements for engaging in means-end reasoning or planning. In what follows I criticise all three of these commitments. First, I undermine Sterelny's claim regarding the transparency of internal environments. I argue, below, that his defence of the claim is insufficient, and consequently that internal environments can also favour robust tracking. Second, I argue that even if internal environments were transparent, it would not follow that cue-based control processes would be generally sufficient. I call the condition in which cue-based control is sufficient 'motivational transparency', analogous to informational transparency. In the following section, I argue that motivational transparency does not generally follow from internal transparency. Finally, I maintain that Sterelny has made an unsatisfactory choice of explanatory target in his discussion of preference. I argue in the subsequent section that rather than focusing on means-end reasoning, or represented goal states, what is needed, and prior to either, is a notion of incentive values, attached to occurrent environmental information and to possible actions.

The lines of critical thinking offered here are not exhaustive, and the arguments I provide are brief. I aim to highlight some difficulties with what Sterelny himself identifies as 'tentative' moves in the area, with the aim of advancing the same general project. I will not have space to develop or defend a positive view distinct from Sterelny's, although some hints will emerge.

Limits on Internal Transparency

As explained above, Sterelny maintains that internal environments, having homogenous interests, will be devoid of hostility, and so be under pressure to develop accurate and transparent signals of biological states and needs. The interests within an organism are presumably homogenous (Sterelny does not spell this out explicitly) because all parts of a single organism are in some sense equal shareholders in whatever reproductive success the individual organism enjoys.

This suggestion plausibly applies, subject to cost constraints, to internal signals, to the extent that internal interests coincide. This they *mostly* do. The relative absence of hostility does indeed imply the internal absence of one source of pollution in the external informational environment.

But hostility is not the *only* source of departures from transparency. As Sterelny says, an environment is *informationally translucent* when states that matter to it 'map in complex, one to many ways onto the cues they can detect' (Sterelny 2003: p. 21). These conditions can, and do, arise in hostility-free internal environments, in a number of different ways. Here I identify three considerations:

Limits on transduction: Just as in the external case, not all internal states have unique signatures that cost-effective transducers can specialise in detecting. Non-nutritive sweeteners, for example, trigger transducers whose proper function is to respond to sugars that *can* be digested. The responses of salt receptors, depending on the action of ion channels, are also sensitive to the ambient sodium concentration in the organism, so the resulting neural signals can be highly ambiguous (e.g. Bertino et al. 1982). When there are relevant limits on transduction, then the internal harmony of interests will not lead to informational transparency.

Complex mappings: Motivationally relevant states can also depend on multiple cues. Information about temperature in humans, for example, is drawn from multiple receptors of different types that are distributed non-uniformly across the surface of the body. Thermoreceptors partly illustrate the previous problem of limits on transduction. They do not come in a single type detecting 'objective temperature'. Instead, information about temperature depends on combinations of receptors for cold and heat, as well as additional nociceptive receptors for extremes of each. But thermoreception also exemplifies complex mappings. As Akins notes, even on the human face the ratio of cold to warm receptors varies from about 8:1 on the nose, to 4:1 on the cheeks and chin, while the lips have almost no cold receptors (Akins 1996: p. 346). Any 'net' signal that might drive behaviour will require these signals to be integrated in some way. More generally, internal states can span multiple organs and tissue types, with varying speeds of signalling and latencies in responding to actions that affect them.

Cost versus accuracy tradeoffs: There are costs to improvements in tracking, just as in the external case. Simply adding internal transducers increases information load, along with metabolic and other costs of building and running the receptors. Psychophysical processes generally do not try to track objective magnitudes, but rather compress transduced variation into a baseline-dependent encoding, where the baseline itself is variable (Barlow 1961). In addition, the further a body is from being a dimensionless point, the more internal signals will tend, sometimes, to be distal, delayed, or both, and subject to the typical error types that arise from distal signals. (One such error type is a false positive from stimulation on any 'labelled line' channel between transducer and brain. Cutting off a finger does not automatically update the topographic neural maps of the body, nor does it deactivate nerves from hand to brain.)

We should conclude that even though internal environments are not generally *hostile*, they can certainly be *translucent.* And translucency favours robust tracking. So Sterelny's premise is at least not straightforwardly or generally true. What about the inference that he draws from it?

Internal transparency does not imply Motivational Transparency

In the previous section I argued that there could be benefits to robust tracking in internal environments because (just as with external ones) there are limits to transparency. Since Sterelny argues from internal transparency to the non-generalisability of his treatment of belief, this is a problem for his position. But it is not the only one. To see this, let us assume that internal environments *are* fully transparent, in the sense that the precise level of deviation from homeostasis of all relevant internal variables are signalled in a consistently high-fidelity way. Even then, cue-bound control can be inefficient.

One reason for this is that *needs can have multiple satisfiers*. A cold animal might be able to make itself warmer, inter alia, by shivering, by huddling with conspecifics, or relocating to a warmer spot. A dehydrated animal can drink, or it can eat, since almost all food contains some water. And so forth. A hungry animal might have more than one foraging option. Accurate information about needs does not always, then, suggest a unique 'good enough' response that would favour cue-bound control.

A further reason is that *actions typically have multidimensional costs and benefits*. As noted, most eating rehydrates as well as nourishes. Different opportunities to eat, or drink, have their own costs in energy, time, extent of competition for the same resource, etc., and their own risks including predation en route, or at the site itself, as well as payoffs in quality and quantity of the resource itself. Costs and benefits can have sharply varying fitness implications—being a little tired or hungry quite frequently is nowhere near as bad as being eaten even once. Even if an animal had accurate information about *all* of these contingencies, it would not generally be obvious what course of action was appropriate or efficient. (We already know that it can be difficult to work out what to do in games of perfect information such as chess.) And animals typically *do not* have most of this information, which favours—for at least some of them—being able to sample the environment and be sensitive to the *returns* from various policies.

Sterelny is aware of these considerations, but does not regard them as favouring the development of preferences. An important part of the reason for this is his view that many animals can deal with the problem of trading off different courses of action by means of the 'built-in motivational hierarchy' (p. 81). This seems likely to be true, at least up to a point. Sterelny is surely correct that selection can act on relatively simple pairings of internal needs and behavioural responses, and that a hierarchical repertoire of such responses can sometimes be a satisfactory solution to the challenge of action selection. That this much is so, however, is not a reason for thinking that the arguments (conditionally) favouring decoupled representation do not generalise to motivational states under *any* conditions. As with detection systems, and belief-like states, we should consider the relative strengths and weaknesses of more or less quick and dirty, or inflexible, procedures.

A fixed hierarchy can probably produce quick, and good enough, responses in a wide range of situations. But such brittle solutions have the very problems of inefficiency under conditions of informational translucency that Sterelny explained when focusing his attention on beliefs, including vulnerability to exploitation by other agents (See above). And if the gains in efficiency from a less rigid approach outweigh the costs, then something other than a fixed hierarchy might pay its way. What this something else might be, I suggest, is relatively general (across actions and environmental states) sensitivity to *reward*. Then the specific profile of things to be found rewarding can be set by processes of natural selection, and the organism's behavioural dispositions partly shaped by experience of action-reward relationships. If we combine the argument of the preceding section and this one, we see a case for the robust tracking of motivationally relevant states, *and* a role for motivation in prioritising actions given response breadth. What motivation should do, what it is *for*, is prioritising on the basis of the *returns* from actions.

What I am describing, though, sounds rather different from what Sterelny sets up as his explanatory target. This is deliberate, and in the following sub-section I attempt to justify it.

The wrong target

As noted above, Sterelny takes the target for an account of the motivational part of folk psychology to be representations of goal states, and capacity for means-end reasoning selecting actions that

bring the goal states closer. Recall, though, how he describes his overall project, as relating the 'wiring and connection facts' about human cognitive architecture to the 'interpretation facts' which are the elaborations of our folk self-conception as intentional agents. This means relating wiring and connection facts to beliefs, on the one hand, and desires on the other. Or, perhaps, to one or both of them as elaborated or regimented by some science (or group of them) that takes the folk conception as a starting point. So, we can ask, what is the approximate functional content of the folk notion of desire, or an appropriate scientific regimentation of it, for relating these two kinds of fact?

I propose that the core of the folk conception is a fairly general (and sometimes imprecise) notion of motivational strength. An intentional agent desires different goals, or to perform different actions, to varying degrees. When two mutually exclusive actions are available, it does the one that it wants the most.

Such a loose and general notion is at least prima facie compatible with some of the leading philosophical accounts of desire, even though the field is contested. A leading contender is the view that desires are dispositions to action, given beliefs (e.g. Smith 1987). Teleosemantic theories of desire are dispositional, and also offer an analysis of the biological function of desires (e.g. Millikan 1984; Papineau 1987). One competing approach is provided by theories of desire based on pleasure, for example Morillo (1990), which approach in addition identifies the dopamine system in the brain as the basis of pleasure. Details of this view are disputed by Schroeder (2004), who associates desire with learning, and identifies the dopamine system with reinforcement learning, rather than pleasure. (There are also theories of desire less obviously friendly to naturalists, such as broadly Socratic ones connecting desire to judgements about what is good.) Without joining those disputes, I note that the disposition, pleasure and learning accounts all have weaker commitments than Sterelny's target ('representations of the external world', and capacity for means-end reasoning). Where these theories require representations of the world, those are mostly provided by *beliefs*. What desires do is relate world-states, whether represented or cued by occurrent experience, to tendencies to action. As Papineau puts it, beliefs should be thought of as having 'no effects to call their own', but then which effects are produced depends on the motivational states, i.e. the desires (Papineau 2004: p. 494).

Matters do not change substantially if we shift focus to consider scientific theories as another source of what Sterelny calls 'refined versions of our folk self-conception'. In behavioural psychology and economics (which are among the leading scientific regimentations of something that might be related to the folk concept of desire) the key notions are reward or reinforcement or utility, which are considered to provide an ordering of desirability for states and actions (See Spurrett 2014). In addition, in behavioural neuroscience the drive-based theories that Sterelny seems to favour for organisms in which preferences (as he understands them) have not developed have largely been displaced by incentive-based approaches. Here too, the theories have more modest commitments than Sterelny requires. What preferences fundamentally represent are rewards (or reward expectancies), not states of the external world, even though there might be preferences related to states of the world that the system can represent anyway.

None of this is to say that means-end reasoning is neither interesting nor important. When it occurs, it plausibly stands in need of an evolutionary rationale. From the perspective suggested here, though, means-end reasoning is primarily a *representational* achievement, consisting in the capacity to simulate transitions between world-states, including transitions occasioned by actions, and evaluate them using the same general sensitivity to incentives as apply in 'on-line' experience (see Shea et al. 2008). Sterelny is probably correct, furthermore, that means-end reasoning is relatively incompletely developed even in humans, and only found marginally in relatively few nonhuman animals.

Conclusion

According to the Environmental Complexity Thesis (ECT) 'the function of cognition is to enable the agent to deal with environmental complexity' (Godfrey-Smith 1996, 2002). Sterelny (2003) develops an account of folk psychology within the general terms of the ECH. He argues that belief-like states can be explained as a response to failures of environmental transparency, combining robust

tracking (sensitivity to multiple types of detectable information) and response breadth (the relevance of registered states to more than one behaviour). But, he argues, in the case of motivation internal environments will tend to be transparent, and because of this the inference from translucency to (an approximation of) a folk-psychological kind does not apply. Preferences, understood as capacity for means-end reasoning about representations of the external world, are not predicted, at least for most organisms, because transparent signals of internal state plus a built-in hierarchy of drives are a pretty good way of prioritising actions.

I have accepted the Environmental Complexity Thesis, and broadly support Sterelny's treatment of belief-like states, but argued against significant parts of his treatments of desire-like states. Internal environments are not as transparent as he thinks, with the result that there is work for robust tracking there too. Motivational transparency does not follow from informational transparency either, and so there is work for relatively generalised sensitivity to reward. The view of preference that is predicted here is different from what Sterelny sets out to find, but I have also argued that means-end reasoning is not the most important feature of desire. Preferences are representations of a sort, but they represent the returns (experienced or anticipated) from experienced states, or from possible actions. Most of the burden of argument has been on the negative project of blocking Sterelny's 'no generalisation' argument. The fuller development of the positive picture suggested here is a task for another occasion.

References

Akins, K. 1996. 'Of sensory systems and the "aboutness" of mental states'. *Journal of Philosophy* 93(7), pp. 337–372.

Barlow, H.B. 1961. The coding of sensory messages. In: Thorpe and Zangwill (eds), *Current Problems in Animal Behaviour*. New York: Cambridge University Press, pp. 330–360.

Berridge, K.C. 2004. Motivation concepts in behavioral neuroscience, *Physiology and Behavior* 81, pp. 179–209.

Bertino, M., Beauchamp, G.K., Engelman, K. 1982. 'Long-term reduction in dietary sodium alters the taste of salt'. *American Journal of Clinical Nutrition* 36, pp. 1134–1144.

Brooks, R.A. 1991. 'Intelligence without representation' *Artificial Intelligence* 47, pp. 139–159.

Burt, A., Trivers, R. 2006. *Genes in Conflict: The biology of selfish genetic elements*, Cambridge, Massachusetts: Bellknap/Harvard University Press.

Dennett, D. 1987. *The Intentional Stance*. Cambridge, Massachusetts: MIT Press.

Dickinson, A., Balleine, B. 2000. Causal cognition and goal-directed action. In: Heyes, C. and Huber, L. (eds), *The evolution of cognition*. Cambridge, Massachusetts: MIT Press, pp. 185–204.

Christensen, W. 2010. 'The Decoupled Representation Theory of the Evolution of Cognition: A Critical Assessment'. *British Journal for the Philosophy of Science* 61, pp. 361–405.

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

Godfrey-Smith, P. 2002. Environmental Complexity and the Evolution of Cognition. In: Sternberg, R. and Kaufman, J. (eds), *The Evolution of Intelligence*. Mahwah: Lawrence Erlbaum, pp. 233–249.

Godfrey-Smith, P. 2014. *Philosophy of Biology*. Princeton: Princeton University Press.

Haig, D. 2002. *Genomic Imprinting and Kinship*. New Brunswick: Rutgers University Press.

Haig, D. 2006. Intrapersonal conflict. In: Jones, M.K. and Fabian, A.C. (eds), *Conflict*. Cambridge: Cambridge University Press, pp. 8–22.

Libersat, F. 1994. 'The dorsal giant interneurons mediate evasive behavior in flying cockroaches'. *Journal of Experimental Biology* 197, pp. 405–411.

Millikan, R. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, Massachusetts: MIT Press.

Millikan, R. 1995. 'Pushmi-pullyu representations'. *Philosophical Perspectives* 9, pp. 185–200.

Morillo, C. 1990. 'The reward event and motivation'. *Journal of Philosophy* 87, pp. 169–186.

Papineau, D. 1987. *Reality and Representation*. New York: Basil Blackwell.

Papineau, D. 2004. 'Friendly thoughts on the evolution of cognition'. *Australasian Journal of Philosophy* 82(3), pp. 491–502.
Schulz, A. 2011. 'The adaptive importance of cognitive efficiency: an alternative theory of why we have beliefs and desires'. *Biology and Philosophy* 26, pp. 31–50.
Schulz, A. 2013. 'The benefits of rule following: A new account of the evolution of desires'. *Studies in History and Philosophy of Biological and Biomedical Sciences* 44, pp. 595–603.
Schroeder, T. 2004. *Three Faces of Desire*. New York: Oxford University Press.
Schroeder, T. 2014. Desire. In: Zalta, N. (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition). Available at: http://plato.stanford.edu/archives/spr2014/entries/desire/
Shea, N., Krug, K. and Tobler, P.N. 2008. 'Conceptual representations in goal-directed decision-making'. *Cognitive, Affective, & Behavioral Neuroscience* 8(4), pp. 418–428.
Shea, N. 2014. 'Reward Prediction Error Signals are Meta-Representational'. *Noûs* 48(2), pp. 314–341.
Smith, M. 1987. 'The Humean Theory of Motivation'. *Mind* 96, pp. 36–61.
Spurrett, D. 2014. 'Philosophers Should Be Interested in "Common Currency" Claims in the Cognitive and Behavioural Sciences'. *South African Journal of Philosophy* 33(2), pp. 211–221.
Sterelny, K. 2003. *Thought in a Hostile World*. Oxford: Blackwell.
Sutton, R.S. and Barto, A.G. 1998. *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: MIT Press.