Biol Philos (2008) 23:87–100
DOI 10.1007/s10539-007-9077-7

# Beyond persons: extending the personal/subpersonal distinction to non-rational animals and artificial agents

**Manuel de Pinedo-Garcia · Jason Noble**

Received: 31 January 2007 / Accepted: 2 June 2007 / Published online: 26 June 2007


**Abstract** The distinction between personal level explanations and subpersonal ones has been subject to much debate in philosophy. We understand it as one between explanations that focus on an agent's interaction with its environment, and explanations that focus on the physical or computational enabling conditions of such an interaction. The distinction, understood this way, is necessary for a complete account of any agent, rational or not, biological or artificial. In particular, we review some recent research in Artificial Life that pretends to do completely without the distinction, while using agent-centred concepts all the way. It is argued that the rejection of agent level explanations in favour of mechanistic ones is due to an unmotivated need to choose among representationalism and eliminativism. The dilemma is a false one if the possibility of a radical form of externalism is considered.

**Keywords** Agents · Artificial life · Category errors · Externalism · Eliminativism · Levels of explanation · Mechanism · Philosophy of mind · Representationalism

## Agents and their parts

Philosophers of mind and of cognitive science have tended to blur the distinction between content at the personal and at the subpersonal level. For instance, Dennett (1978) argues that the content of intentional states, such as believing that Tokyo is the capital of Japan or

M. de Pinedo-Garcia (✉)
Departamento de Filosofía, Programa 'I3' (Ministerio de Educación y Ciencia) Facultad de Filosofía y Letras, Edificio B, Universidad de Granada, 18011 Granada, Spain
e-mail: pinedo@ugr.es

J. Noble
School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
e-mail: J.Noble@soton.ac.uk

hoping that it won't rain tomorrow, is a subset of the content of subpersonal computational states of the organism. Likewise, some important scientific research on subpersonal[1] states that enable cognition has tended to completely ignore such a distinction. Most famously, David Marr's impressive work on vision is tainted by the author's tendency to move from the issue of the non-conceptual content of computational operations on sensory input to the issue of the conceptual content of perceptual, organismic states. A great deal of research in contemporary neuroscience follows Marr in this regard. The main difficulty with this line is that subpersonal states seem to have content only inasmuch as they facilitate agent/environment interaction. ''Genuine'' content is to be ascribed only to the whole organism: only the organism has states that can be described as propositional statements about its environment. Amongst the problems that inattention to the divide brings, perhaps the biggest is confusing the normative and externalist requirements of agent-content with the merely causal requirements of subpersonal content, a confusion that echoes the naturalistic fallacy. We would like to argue that such a fallacy is a burden that a proper naturalism can do without.

Moving uncritically from the agent to the subagent level can lead to unwittingly circular argumentation: clearly, an agent can take in features of its environment in virtue of possessing a certain complex internal structure; however, to suppose that such a structure is all there is to intentionality (i.e., to having states directed towards contents) threatens to cut off the link between the inner and the outer and to make a mystery of the ability to perceive and act on the world. The manoeuvre is simple: if agent-level content is made possible by computational states, we are entitled to ascribe content to such states, even if only metaphorically. Given that such states are easier to deal with than agent-level states, it is tempting to argue that a satisfactory explanation of the former would give us all we need to understand the latter.[2] However, we can only ascribe content to subpersonal states inasmuch as they are enabling conditions for personal ones: our subpersonal explanation inherits its contentfulness from the broader explanatory project of accounting for the agent/environment pair. It would be disingenuous to forget that the content ascribed in the first case is an abstraction from content that can only be found in agents.

The need to reconcile these two different explanatory projects, one dealing with situated agents considered as a whole, the other with the mechanisms that make possible such situatedness, is not limited to the case of rational, language-using animals or even to non-linguistic animals in general. That would be the case if the distinction between both projects relied on the idea that the rational features of persons cannot be accounted for in purely mechanistic terms or on the idea that there is something unique about naturally occurring animals, rational or not, that demands a special kind of explanation. It is central to our argument, however, that a parallel need to reconcile two explanatory projects is found in the domain of artificial life (ALife) where the agents involved are currently neither rational nor language-using and exist only in simulation. The simulated agents of ALife research can be explained in terms of their mechanical, subpersonal properties (e.g., the specific pattern of connection weights in an evolved neural network) or in terms of

[1] The reference to persons in ''personal'' and ''subpersonal'' should be read as accidental. We use the adjectives to refer generally to agents and to their internal machinery.

[2] This is, in fact, the position we find in Dennett's paper mentioned above. His take on these issues is far more complex and oscillating than this. For instance, we can observe a move from blunt intrumentalism (1971) to some form of transcendentalism regarding the presumption of rationality for agents (1981) and back again to instrumentalism with realistic undertones (1991). Each of these accounts of intentionality are worth studying on their own. However, the position defended in his paper from 1978 is sufficiently popular to deserve discussion independently of the subtleties of Dennett's work.

higher-level properties that take into account the agent, its environment, and their historical interaction (e.g., the effects of a particular history of selection in a given environment). ALife researchers are aware of this split, and often implicitly or explicitly endorse explanatory schemes from biology, such as the ethologist Tinbergen's (1963) distinction between explanation in terms of mechanisms and explanation in terms of evolved function or adaptive value.

To describe the function of something is one way of explaining it: one might explain a can-opener to a curious child by saying that its function was to open cans. However, functional explanations are of course not the only explanations that can be offered of animal behaviour. Tinbergen suggested that there were four distinct classes of question that could be asked about an animal's morphology or behaviour: questions of mechanism, ontogeny, phylogeny and function. Mechanism refers to the way the behaviour or trait is physically realized; in other words, how does it work? Ontogeny and phylogeny are historical questions and account for traits by appeal to their developmental and evolutionary history respectively. Questions about function refer to the selective advantage that a trait has for the animal; the question has also been described as one of ''survival value'', ''adaptive value'' or of ''ultimate function''.

In our proposal regarding the need for agent-level concepts to make sense not just of complex animals but even of artificially evolved virtual agents we will make use of considerations of an externalist nature and opt for a constitutive rather than merely causal understanding of the role of the environment. In other words, we will oppose the eliminativism that would see agent-level concepts discarded in favour of causal/mechanistic explanations involving causal relations between the agent's insides and its environment. Instead we argue that such causal relations could only be grasped once the relevance of the environment for the agent has been properly understood; against a representationalist reading of agent-level concepts we will claim that no understanding of an agent can be complete without incorporating in its description (rather than as some sort of appendix) the relevant features of the environment. Tinbergen's defence of the need for functional explanations belongs in the same framework as our argument. However, his plea for a pluralistic approach was based on an appeal to the extant work in biology and ethology in 1963 and to linguistic usage within the scientific community. We aim at grounding the distinction with philosophical argument. We feel that such an argument does not need to rely on considerations of rationality (which would preclude extending the distinction to agents that are not themselves consumers and providers of reasons) nor to have a non-naturalistic nature. Tinbergen legitimizes functional questions in ethology mainly by showing the interesting and productive results that have been obtained by asking them. He is not trying to anticipate a future where philosophers or even scientists inspired by the development of responses to mechanistic questions (or to ontogenetic and phylogenetic questions) would claim that those are the only legitimate ones. Furthermore, Tinbergen does not limit functional explanations to agents but is happy to move further down (e.g., to the level of organs, cells, or specific behaviours) in their defence. His split is not exactly the same as ours, but it is a productive starting point for our argument. In short, our purpose is to show that Tinbergen was right in his defence of functional explanations, to provide a philosophical argument supporting that defence, and to apply it to the field of ALife.

For our own purposes, there is an excellent match between mechanistic explanation in biology or ALife and subpersonal explanations of cognition. However, we will need to look closely at just how deep or shallow the parallel is between personal-level explanation, which involves reasons for action, and an explanation that appeals to evolved function, which is ultimately a historical story about why one trait won out over another in a history

of selection. Certainly Millikan (1984) argues that evolutionary explanation can solve the normativity problem regarding content, but we are not convinced that personal-level explanations can be done away with in this way.

ALife research covers a broad territory, but it can reasonably be summarized as seeking to offer an understanding of life and mind through synthetic means. Some work in ALife seeks low-level mechanistic explanations of agent behaviour, while other researchers consider their simulated agents in terms of ecology and function, and are agnostic about mechanisms. Some of the best research is carried out in an attempt to integrate both levels of explanation, e.g., Randall Beer's work on ''minimally cognitive agents'' (see Beer 1990, 1996). Beer examines mechanisms, but also has a constant concern for situating the agent in its environment. This attention to an agent's context is reminiscent of the cybernetic movement of the 1950s, and also of J. J. Gibson's ecological psychology. Work like Beer's may even provide a template for explanation of much more complex systems such as human beings. However, not all ALife researchers are so careful, and we plan to identify problematic slippage between the two levels of explanation, usually resulting from an identification of the agent's perspective with a component of its architecture.

One of the reasons for ALife research to have focussed on this sort of mechanistic explanation is because of the way in which the field has defined itself in opposition to the classical artificial intelligence tradition (often referred to somewhat disparagingly as ''GOFAI'', for ''good old-fashioned artificial intelligence''). In classical AI, an agent's behaviour is explained as the result of it planning a course of action based on its internal model of the world. The agent is supposed to be using sensory input to update its representation of what is going on in the world, and then to be manipulating these internal representations in order to plan and re-plan the optimal way to achieve its explicit goals. ALife evangelists such as Brooks (1991) and Harvey (1996) were quick to point out problems with this picture, and to supply counter-examples in the form of simple reactive agents that were capable of highly competent behaviour without possessing anything resembling an internal representation. If you could completely describe the workings of an agent's neural network in simple mechanistic terms, as done by Cliff et al. (1993) for example, you could then ask, ''Where are the internal representations?'' Nowhere, of course. An important paper by Beer that will serve as a key example for us is also driven by this ongoing rejection of representationalism: Beer argues that dynamical systems explanations (i.e., mechanistic explanations) raise ''important questions about the very necessity of notions of representation and computation in cognitive theorizing'' (Beer 2003: 210). We will argue that the move from rejecting representations to rejecting all forms of intentional description of an agent's behaviour is unwarranted.

Readers may object that Beer's and other ALife agents do not have a real environment, they only have a simulated environment. On this we agree. We are not trying to dislodge real agents and real environments from their privileged position as the ultimate objects of our explanatory efforts. However, we place no stock in essentialist arguments that there is something special about a particular real environment that no simulated environment could possibly capture. We see the importance of the environment as being simply the sum total of the ways in which it influences and is influenced by the agent, and we maintain that although it would be a huge technical challenge, there is no in-principle reason why the relevant features of a given organism's environment could not be effectively simulated. In the case of Beer's thought experiment, the environment is many orders of magnitude less complex than even that of the simplest real animals. However, that does not prevent the experiment telling us something useful about cognition: about the ways in which

environments and agents may interact, and thus about the proper grounds for explaining the behaviour of agents.

The worry about simulation could proceed as follows, however: by confusing the representation with the thing represented we are already assuming what we want to defend; in order to understand the behaviour of a real animal we need to make use of agent-level concepts and to appeal to features of the (real) environment that are salient to the agent. A simulacrum of an agent will also demand those kinds of concepts and appeals inasmuch as it is modelled on agents. However, as it will become clear in our discussion of Beer's paper, we borrow the author's quasi-realistic approach to his own thought experiment in order to argue for a philosophical conclusion which diametrically opposes his. Our claim will be that the only way to understand the analysis presented in his paper is to take seriously the casual overview of the experiment: a minimally cognitive agent that distinguishes diamonds and circles.

Another potential challenge that ALife poses for the personal/subpersonal distinction is exemplified by Braitenberg's 1984 book *Vehicles: Experiments in Synthetic Psychology* (a work that predates the ALife movement but that has formatively influenced it). In his fable on the evolution of intelligence, Braitenberg shows that synthetic agents can appear to warrant a personal-level description (or at least an anthropomorphic description) despite having an extremely simple internal structure. This is a challenge to the personal/subpersonal distinction because it suggests that sufficiently simple agents may, despite appearances, only warrant mechanistic explanations for their behaviour. This in turn could raise questions about whether *any* agents require personal-level explanation. We plan to answer this challenge in two ways: by showing that even the simplest agents deserve ecological and functional explanations, and by demonstrating that such explanations must eventually be replaced by personal-level explanation as the relevant system becomes more complex.

## Levels of description in the explanation of behaviour

When facing the task of explaining the behaviour of an autonomous agent, be it a person, an animal or an artificially evolved agent, the focus can be on the internal machinery that enables the agent to act, on the agent as a whole, or on the coupled system consisting of the agent and its environment. There is a long philosophical tradition behind the idea that none of these approaches can be reduced to any of the others.[3] However this tradition tends to argue against the reducibility of agent level explanations to mechanistic explanations (explanations in terms of the agent's internal mechanisms) *in the case of persons*, and such arguments often highlight issues dealing with the normativity of language and socialization. Our argument extends the distinction to all organisms and agents capable of inhabiting their environment competently without appeal to considerations of rationality. We sympathize with the idea that understanding rational action by appeal to nomological explanations (functional, physical, dispositional, etc.) of the relationship between a person's nervous system and naturalistically describable features of her environment constitutes a category error. However, the core of our point is something that we take to be close to a truism: personal (agent) level concepts are necessary to isolate the relevant features of the environment that mechanistic, computational or dynamical systems

[3] The tradition we are referring to includes Wittgenstein (1953), Ryle (1949), Davidson (1970) and Dennett (1978).

explanations need for their tasks and, hence, any reduction of the former concepts to the latter explanations would be circular. We will illustrate such a need by centering on some projects within the field of ALife which shed light on the distinction and, in particular, by means of the analysis of a very simple agent, presented by Beer (2003) in a recent issue of *Adaptive Behavior.*

A typical ALife paper introduces an evolved agent that does something, and then the body of the paper is dedicated to explaining how it does it. For instance, in the case we will analyze, Beer evolved a circle-catching, diamond-avoiding agent and spent a heroic amount of time studying the coupled dynamical system of the agent and its environment in an effort to make sense of the agent's ability to catch circles and avoid diamonds. The temptation of thinking that ALife research consists exclusively of what the body of one of these typical papers offers, i.e., mechanistic explanations (in the case of Beer, the dynamical systems account that explains the agent's abilities without needing to invoke internal representations), neglects the fact that such explanations are only interesting inasmuch as they shed light on broader questions about life, cognition, and the ways in which agents are situated in their environments. Some ALife papers are more ambitious than others in that they try to alter our understanding of life, agency or cognition without aiming to dispel one of these levels of explanation, or reduce one to the other. Authors like Maturana and Varela (1980) may want us to alter our conception of what sort of agent-level explanations we should be seeking, but do not want us to abandon that project entirely.

Our argument is twofold: agent level explanations cannot be reduced to subagent ones and, furthermore, subagent explanations make ineliminable use of agent level concepts. We are committed to the idea that the agent/subagent distinction is real and that the personal/subpersonal distinction is one of its subclasses. By offering an argument for the reality of the former distinction we hope to illuminate some foundational issues for the cognitive sciences, as well as to offer new reasons for defending the latter.

## From personal/subpersonal to 'personal'/'subpersonal'

Gilbert Ryle's seminal *The Concept of Mind* is sometimes remembered as an original rejection of Cartesian immaterialism and as an alternative theory about the mind, incorporating a materialism akin to behaviourism in psychology. However, Ryle's rejection of the myth of the ''ghost in the machine'' is both a rejection of the ghost and a rejection of the machine (i.e., a rejection of immaterialism and of mechanism). It is not that the mind is a mechanism, on a par with those that make up our nervous system. The truly deep influence of Descartes in contemporary thought about the mind is not his immaterialism, but rather the idea that the mind is an object, subject to the same kind of scientific explanations as any other object. These explanations are understood as going from observable properties to hidden causes. In the case of the mind, in the neo-Cartesian reading, overt behaviour is supposed to be explained in terms of, say, internal information-processing devices. Using the distinction between personal and subpersonal styles of explanation, Ryle's point, regarding persons, is that often the explanation of actions stops quite early, with an appeal to reasons. In other words, rational explanations are normative in a double sense; not only do reasons belong in the realm of justification rather than in the realm of causal, nomological, explanations, but furthermore we do not need to look for hidden causes to fully understand normal behaviour: ''The classification and diagnosis of exhibitions of our mental impotences require specialized research methods. The

explanation of the exhibitions of our mental competences often requires nothing but ordinary good sense (...)'' (Ryle 1949: 308). When we get something right, our behaviour is best explained at the personal level, in the ordinary language of beliefs and desires: John drove to the supermarket because he wanted milk and thought that it was available there. On the other hand, when we get something wrong, a sub-personal explanation, phrased in terms of physical interactions between our component parts, may be called for. Suppose John crashes his car on the way to the supermarket, because he suffers a mild stroke. The relevant explanation is obviously sub-personal.

Once we have granted that John is a rational agent, there is not much more to say about his (successful) trip to the supermarket. Ryle was content with this short chain of reasons for rational acts, and was extremely sceptical about the idea that psychology or any other science could somehow supplant personal-level explanations and supply the 'real' causal story behind a person's actions. Sub-personal explanation, on the other hand, can go very deep: accounting for John's stroke might involve considerations of diet, physiology, genetics, biochemistry, and ultimately physics. Sub-personal explanation is also part of the story in explaining John's competencies: how is it that he has the sensorimotor coordination needed to drive a car, or to remember the way to the supermarket? It was in these sorts of questions that Ryle saw a role for cognitive science.

Furthermore, Ryle held that the category errors generated by confusion between the two different levels of explanation were responsible for most of the apparent mysteries about cognition. Ryle's most famous example of a category error involves a visitor being shown all of the buildings in Oxford and then insisting ''Yes, but where is the university?'' A similar error is easily committed when thinking about agents and their component parts. For example, John is like the university: he is in a sense constituted by his component parts, but he is not the same type of thing as one of his components. Forgetting this, and imagining that the components of John's brain and John the person are on a par with each other, leads to conceptual disasters such as the 'mystery' of how John could possibly be conscious, for example. (If John is equated with the mere matter of his brain then certainly there is a mystery about how the-matter-that-is-John achieves conscious awareness, but if we recognize that John and his brain matter are concepts at different levels, then wondering about how John could possibly be conscious is properly recognized as being a bit like asking how it is that a university can make decisions despite being constructed from stone.)

Ryle's vindication of common sense, as well as his appeal to a distinction between normal and abnormal behaviour, are in line with most defences of the personal/subpersonal distinction (for instance, Davidson's (1970) influential arguments in favour of the holism of the mind and the need to assume rationality in order to understand the actions of linguistic creatures). Personal-level questions are often fully answered when the subject's reasons for her actions are given. However, stopping at that could lead to conflating two different, and equally important, distinctions. On the one hand, Ryle and Davidson reject explanations of phenomena that involve rationality in terms of parts of the rational subject. By doing so, they call our attention to the difference between rational and mechanistic explanations, a distinction that may well be limited to human beings (or rational beings in general). In this sense, the kind of reason-giving explanation suitable for rational agents could be of no use for non-rational agents and one might be tempted to equate parts of persons with agents that are not persons (e.g., non-rational animals, or artificially evolved agents) and to claim that mechanistic explanations are all that is needed both for agents' insides and for non-rational agents. In parallel with the distinction between personal and subpersonal explanations for persons, it is possible to argue for the need to distinguish

between an agent level explanation of non-personal agents (i.e., agents that are not persons) and a subagent explanation in terms of their parts.

A notable exception to this concentration on rationality in the defence of a personal/subpersonal distinction is McDowell (1994). McDowell, while having dedicated most of his work precisely to issues dealing with rationality, has also argued that the relevance of the distinction is not limited to rational animals, but extends to any creature capable of ''competently inhabiting its environment''. McDowell employs the example of predatory behaviour in a frog, which we might seek to explain both in terms of the neurological parts of the frog and in terms of the frog as a complete agent. He draws on the landmark neuroscience paper ''What the frog's eye tells the frog's brain'' (Lettvin et al. 1959) and suggests that to fully understand the frog, we would also need a ''froggy/sub-froggy'' distinction. As well as explaining the frog in terms of its neurological components, we need to consider the frog as a whole agent in its environment or *Umwelt*, and look for example at the significance of different environmental features *for the frog*. This corresponds to what we have called agent-level or ecological explanation. He argues that any account of, say, perceptual awareness of features of the environment involves essential use of concepts suitable to those features.

> How, then, does the *frog* get into the act? I suspect that this question (...) tends to be suppressed because of an unfortunate feature of the otherwise excellent distinction between the personal and the subpersonal. Theories of internal information-processing in frogs are at best ''sub-personal'' (...), not sub-personal, because there are no persons around in contrast with whom we can mark the standard distinction. (...) The point of saying that the theory of internal information-processing in frogs is ''sub-personal'' is not that no persons are involved (...) but that the fundamental idea of such a theory is the idea of informational transactions between *parts* of frogs (McDowell 1994: 347).

We want to go one step further. Any explanation of an agent's behaviour, natural or artificial, must include externalist concepts. Furthermore, those concepts will apply to features of the environment that are salient for the agents, rather than to pieces of information that the agent's insides successfully compute over.

## Looking over the fence to ALife

McDowell, in the paper we have quoted from, compares two highly influential approaches to visual perception, Marr's (1982) and Gibson's (1979). A common conception of the difference between both approaches, one encouraged by Marr himself, sees Gibson's ecological theory as an insightful but incomplete account of perceptual experience, and blames the incompleteness on Gibson's lack of a proper theory of the computational goings-on inside the perceiving creature. Marr, according to this take, provides us with what Gibson misses. However, McDowell argues, to see matters this way is to misunderstand Gibson's project completely. Gibson's aim is not to search for the neurological or computational enabling conditions of perception (a valid scientific enterprise brilliantly explored by Marr's groundbreaking work), but rather to give sense to the essential situatedness of agents in their environment. No purely computational theory can account for the agent's ability to 'see something as ...' (as, e.g., a predator, food, drinkable, etc.). To explain what we mean we will focus on what, to our knowledge, is the most completely

worked-out artificially evolved agent in the literature, Randall Beer's circle-catching device (Beer 2003).

Beer's agent inhabits a two-dimensional world. It can move from left to right and back again along a fixed track. The agent is equipped with a fan-like array of seven range sensors, pointing upwards and subtending a 30-degree angle. The range sensors return no signal if they detect nothing within their maximum range, and an increasing signal as an object is detected closer to the agent. Internally, the agents are equipped with a five-neuron continuous-time recurrent neural network. Signals from the range sensors are fully connected to the five interneurons, with bilateral symmetry enforced. The interneurons are in turn connected to two motor output neurons, which can be thought of as operating two thrusters, one pushing the agent left and the other right. The physics of the simulated world do not incorporate friction or inertia; the agent simply moves at each time step according to the relative activation of the left and right motor outputs.

Periodically, circles and diamonds appear in the ''sky'' above the agent, and drift downwards at a constant velocity. These falling shapes present an environmental challenge for the agent, as circles are good for the agent (think of them as food) and diamonds are bad. Ideally the agent should use its minimal sensory system to figure out which shape is falling, and then catch circles, i.e., be in a position to intercept the circle when it arrives at ground level, and avoid diamonds, i.e., be anywhere else except under a diamond when it hits the ground.

Life for the agent consists of a series of trials in which circles and diamonds are dropped from a constant height but different lateral offsets from the agent's position. Each agent goes through life alone (i.e., there is no direct competition for circles) but is nevertheless involved in an evolutionary process. Agents with the highest success rate at catching circles and avoiding diamonds are more likely to pass on their genes (essentially the configuration of their neural networks) to the next generation. Across many generations of evolution, agents become better at getting things right, i.e., catching the circles and avoiding the diamonds. By the end of the simulated evolutionary process, the best evolved agent has a success rate of better than 97%.

Most of the length of Beer's (2003) paper is dedicated to a painstaking mathematical and statistical analysis of this minimally cognitive task. The author uses this work to launch an attack on representational theories of cognition (in that his analysis makes no use of specific circle or diamond detectors inside the agent) but does not stop there: given that no internal representations need to be postulated (i.e., no internal concepts such as ''diamond'' or ''circle'' are necessary to explain the agent's behaviour) we can also do without any appeal to worldly properties such as 'being a diamond' or 'being a circle'. Neither the concepts nor the properties appear in the explanation of the behaviour. However, they do appear in the description of the agent: it is the amazing ability of the simple agent to differentiate between the two kinds of shape that is itself analyzed in such detail![4] This would be too obvious to be worth pointing out, if cognitive scientists and philosophers of eliminativist tendencies did not hurry to conclude that all agent level talk can be reduced to, or replaced by, computational talk. Such a conclusion is motivated by thinking that the

[4] It could be argued that even the simplest designed artifact or program (say, a thermostat) would also demand both agent-level concepts and an appeal to macroproperties to explain its behaviour. In this paper we are only committed to the idea that it is a sufficient condition for treating something as an agent (and, in parallel, to defend the need for concepts making reference to features of the environment that are relevant to the agent as a whole) that the thing in question is the result of an adaptive evolutionary process. Whether this is also a necessary condition for agenthood is a question that we would like to leave open. We will briefly return to it in our discussion of ecological explanations of animal behaviour.

following dilemma is exhaustive: knowing, and in general being a subject of cognitive processes, is traditionally understood in terms of possession of a complex, internal, representational universe that allows for interesting interaction, through action and perception, with the environment. But there is no way to account for such interaction without studying the whole brain/body/environment dynamical system. Hence, there is no need for representations nor for meaning, or knowledge or cognition to be brought into the picture to understand an agent.

Representations, being internal, cannot explain perception and action, because these are properties of the coupled agent/environment system and not of the internal machinery of the agent. Hence, we do not need to talk of cognition; internal state is enough, but there are no representations or meanings in internal state. The choice is, therefore, either representationalism or behaviourism with a relevantly complex nervous system. But wait a minute. Who said that the only way to understand cognition, knowledge and meaning is in terms of (syntactic) manipulation of internal representations? It is correct, and laudable, to insist that once representations are allowed fluid redefinition and become seen as features of the coupled system there is no need for keeping the label ''representation'', on pain of confusion. But it is not correct to concede to the representationalist that cognition and so on must be internal to the agent. Accepting so is precisely accepting the main motivation for representationalism. After all, the argument for representationalism runs in the following direction: given that thought happens in me (inside my head, my soul, my phenomenological space, my brain, whatever) there must be internal representations for my thinking to play with. Intuitive internalism thus gives rise to representationalism and not the other way around. Therefore representationalism is best dealt with at its source: that is, by challenging the notion that thought is internal.

There is some irony here. In an important sense, ALife is the last place to expect internalism. Unlike artificial intelligence, which smacks of internalism in its resolve to keep mind and world separate, ALife researchers have been quick to appreciate that cognition can be distributed across the agent-environment divide. Witness the popularity of the key phrases ''situatedness'' and ''embeddedness'' in the ALife literature. So we need to be clear about what we are claiming: that the diagnosis of representationalism's flaws has not been radical enough. A last vestige of internalist thinking has resulted in an incorrect association between representationalism and agent-level explanations, and this in turn has produced an unwarranted emphasis on low-level mechanistic explanations of artificial agents.

Why do we feel that agent-level description is of value? Why are we convinced that the impressive analytical vocabulary of dynamical systems theory, for example, is not the only vocabulary needed by the ALife researcher? Again, we refer the reader to the deceptively obvious fact that Beer needs to describe his agent as a circle catcher and a diamond avoider. Indeed, these are the propensities that his agent was selected for over many generations of evolution. This description is admittedly simple, but it is agent-level talk, and clearly of a different explanatory level than a description of the agent/environment system in terms of differential equations. As a quick thought experiment of our own, we ask whether anyone could possibly make sense of the behaviour of the agent given only the dynamical systems description so carefully developed in Beer's paper, and not the brief but enormously helpful agent-level description. Looking only at the mechanistic level, it would be extremely difficult and perhaps impossible to see that all of this complexity was in the service of circle catching and diamond avoidance.

McDowell (1994) makes a similar point in slightly different language. At the agent or ecological level we can pose and answer ''why?'' questions. Why did the frog stick out its

tongue? In order to catch what it saw as a fly. These questions and answers can in turn inspire ''how?'' questions at the mechanistic (''sub-froggy'') level. How did the visual input lead to the appropriate motor output? When we have answered the mechanistic how-question in terms of some sort of neural circuitry diagram, we have described what McDowell calls an enabling condition for the agent-level behaviour. If we were to then insist that this mechanistic explanation could stand alone, we are mistaking an enabling condition for a constitutive one. As Davidson (1973: 247) puts it ''it is one thing for developments in one field to affect changes in a related field, and another thing for knowledge gained in one area to constitute knowledge of another.'' Even the best mechanistic explanation will be incomprehensible without an agent-level framework. If systems as simple as the one analyzed by Beer require on the one hand agent-level explanations and on the other hand a mechanistic description in terms of dynamical systems, then clearly more ambitious targets such as advanced ALife agents, frogs, and human beings will also require both levels of description.

It could be argued that we are just splitting hairs over a terminological matter: we would like to identify cognition with situated agency, while Beer reserves the former label for the alleged internal manipulation of representations that makes agency possible in the first place (and then, by showing that such an account of cognition is unfounded, moves on to identify cognition with brain activity). Maybe so, but there could be more at stake here. Perhaps limiting cognition by locating it within only a part of the whole dynamical system (the brain? the body?) reveals a failure to see the depth of the representationalist mistake. This becomes clear when Beer, following the usage of eliminativist philosophers he approvingly quotes, speaks of ''the mind/brain alone''. Isn't his paper itself a forceful proof that the mind is all over the place; indeed, that the mind is a very abstract way of describing the agent/environment interaction, rather than something internal? Beer is happy to offer such an externalist conception of decisions. Not extending that conception to knowledge and cognition could be a consequence of accepting uncritically some remnants of Cartesian representationalism (after all, it could seem that decisions are more intimately tied to actions than cognition or pieces of knowledge are, but that idea could also be rejected for the higher glory of situatedness). Cognitive science does not need to give up on talking about agent-level concepts such as knowledge and meaning: it just needs to recognize that knowledge and meaning, as much as perception and action, are features of the coupled system and not something internal.

## The relationship between levels of explanation

We have seen Beer's realization that a proper understanding of the sub-agent level, such as the one provided by his mathematical analyses, does not sit well with the traditional view of the agent as operating over internal representations. His conclusion is that we should do away with agent-level talk as such talk cannot be reduced or made compatible with the explanations needed to understand the mechanisms that make possible the interaction between agent and environment. Some authors, such as Bermudez, claim that we need to reject any principled distinction between personal and subpersonal explanations because retaining them invites eliminativist moves such as Beer's. We have suggested that there are independent reasons to rethink the concept of an agent in fully externalist terms. We also feel that such a reconceptualization may be of great benefit for our view of the subpersonal and for our conception of the relationship between both levels of explanation. In this

section we will make use of some proposals by Susan Hurley in her influential book *Consciousness in Action* (1998) which complement the views put forward above. Hurley feels compelled to reject the traditional, Cartesian view of the personal, as a consequence of what we know about the complexity of subpersonal explanation (she does not refer to Beer's work or to work in the field of ALife, but similar considerations can be applied). At the same time, she offers new insight in favour of the distinction that we have been defending.

Hurley blames the dissolution of the distinction on what she calls the ''input/output picture of perception''. According to this picture, perception is exclusively input from world to mind and action is no more than output from mind to world (we would need to replace mind with cognition to apply the criticism to ALife and to the cognitive sciences in general), where the mind (or cognition) is situated in a place of its own, sandwiched between perception and action. The motivation for this picture is a direct projection from the subpersonal to the personal level: if we think of the subpersonal information processing mechanisms that make cognition possible as computing over causal input from the world and producing, as a result, output that causally affects the world, then perception mirrors causal input, cognition mirrors internal computation and action mirrors output to the world (see Hurley 1998, especially pp 288–293). Hurley launches several interconnected objections to this picture: on the one hand, she questions the validity of an internalist characterization of the vehicles of content at the subpersonal level. Instead of thinking about cognition in terms of internal manipulation of information, she thinks of the subpersonal causal flow as a 'complex dynamic feedback system', one that is not linear like the input/output one and that incorporates features of the environment. Such vehicle externalism is very close to the dynamical system approaches we have seen Beer defending and blending with eliminativism. Subpersonal externalism together with constitutive rather than merely instrumental interdependence between input and output is enough to reject a simple projection from causal, subpersonal, explanations to agent-level ones. If input and output depend non-instrumentally on each other, on internal processing and on the environment, the idea of the mind as separated from the world with perception and action as boundaries loses much of its alleged scientific credentials.

On the other hand, if we rethink the subpersonal along the lines of the dynamical systems approach, the subpersonal enabling conditions for perception, cognition and action cannot be seen as discrete (i.e., as coinciding with causal input, information processing and causal output respectively): ''(…) the contents of both perceptions and intentions may in general be carried by the complex dynamic relations between inputs and outputs in such a system'' (Hurley 1998, p. 308; see also pp 245–249).

We would like to finish our discussion of Hurley's proposal with two remarks. First, it is noteworthy that Hurley's approach, unlike the one we have put forward, uses considerations regarding the subpersonal architecture of the agent to support the separation between two explanatory projects. Rather than highlighting features that are salient only for the organism or simulated agent as a whole, Hurley rejects the reduction of agent-level concepts to subpersonal ones by reworking our understanding of the subpersonal level in such a way that it does not invite any clear-cut reduction. This strategy is perfectly compatible with our insistence on the need of agent-level concepts even to make sense of subpersonal explanations or the differential equations used by them. In fact, we have suggested, in our discussion of Beer's agent and of the conclusions that he reaches from the analysis of its behaviour, that a dynamic understanding of the subpersonal is a perfect travelling companion to the externalist, non-eliminativist and non-representationalist conception of cognition that we have offered.

Second, Hurley accompanies her defence of the distinction with a proposal about the relationship between both levels, the 'two-level interdependence view': perception and action are interdependent because their contents depend on complex relations between input and output at the subpersonal level. It would be utterly mysterious if personal level vocabulary, which seems indispensable to so many areas of our understanding of reality, could be done without. It would be no less mysterious that personal and subpersonal level explanations were completely unrelated. We have remained neutral with respect to the details of the links between both enterprises.[5] Hurley's proposal is as good as any to understand the constraints that our research on the subpersonal places on our conception of persons, animals and agents.

Personal level explanations, such as they are understood in this paper, provide the general framework within which subpersonal, causal/mechanistic explanations must be understood. Complementarily, subpersonal explanations offer us the physical details of the mechanisms that enable situated agency. In this sense, personal level explanations are relatively autonomous from implementation details: we can know (broadly) what an agent does in its environment by appealing to its selection history or, in the case of complex organisms, to its reasons to behave, without needing to know everything about the agent, its parts or the environment. In particular, we can do this without needing to know the details of its internal machinery.[6] Conversely, subpersonal explanations only start making sense when we keep in view what the agent is supposed to do or what it actually does, something that is only available from the personal perspective. Agent-level explanations work by highlighting what is salient in the environment for the agent as a whole, rather than by analysing micro-features of the environment or the agent's insides. As we mentioned above, they often finish very soon, but this does not imply that they do not say much. Mary went to the shop because she needed some milk, the birds are flying south because they are migrating to a warmer climate, the agent moved to the left because it was selected to catch circles and there is one falling to its left. Coming back to Tinbergen's framework, they answer 'why' questions, questions without which 'how' questions would not be of much use.

## References

Beer RD (1990) Intelligence as adaptive behavior: an experiment in computational neuroethology. Academic Press, Boston

Beer RD (1996) Toward the evolution of dynamical neural networks for minimally cognitive behavior. In: Maes P, Mataric M, Meyer J-A, Pollack J, Wilson SW (eds) From animals to animats 4: proceedings of the fourth international conference on simulation of adaptive behavior. MIT Press/Bradford Books, Cambridge, MA, pp 421–429

Beer RD (2003) The dynamics of active categorical perception in an evolved agent. Adapt Behav 11:209–244

Bermudez JL (2000) Personal and subpersonal: a difference without a distinction. Philos Explor 2:63–82

Braitenberg V (1984) Vehicles: experiments in synthetic psychology. MA, MIT Press, Cambridge

Brooks RA (1991) Intelligence without representation. Artif Intell 47:139–159

[5] Having said that, we have insisted on the need for agent-centered concepts to make sense of subpersonal explanations and place them in their proper context, and we have suggested that functional and ecological explanations work both at the personal and the subpersonal level.

[6] However, we have endorsed Hurley's idea that a simplistic understanding of the subpersonal, such as the roughly Cartesian input/output picture, may work as a hidden assumption guiding our conception of the personal level. We have praised Beer's profound analysis of the dynamical system that makes possible the cognitive achievement of his agent precisely as acting as an antidote against representationalist reductionism and as being a perfect complement to an ecological and functional view of the agent/environment pair.

Cliff D, Harvey I, Husbands P (1993) Explorations in evolutionary robotics. Adapt Behav 2:73–110
Davidson D (1970) Mental events, in his *Essays on Actions and Events*. Clarendon Press, 1980, Oxford, pp 207–225
Davidson D (1973) The material mind, in his *Essays on Actions and Events*. Clarendon Press, 1980, Oxford, pp 245–259
Dennett DC (1971) ''Intentional Systems'', his *Brainstorms*. Harvester Press, 1978, Sussex, pp 3–22
Dennett D (1978) Toward a cognitive theory of consciousness, in his *Brainstorms*. Harvester Press, 1978, Sussex, pp 149–173
Dennett DC (1981) Making sense of ourselves in his *The Intentional Stance*. The MIT Press, 1987, Cambridge, Mass, pp 83–101
Dennett DC (1991) Real patterns, J Philos 88:27–51
Gibson JJ (1979) The ecological approach to visual perception. Houghton Mifflin, Boston
Harvey I (1996) Untimed and misrepresented: connectionism and the computer metaphor. AISB Quart 96:20–27
Hurley SL (1998) Consciousness in action. Harvard University Press, Cambridge, MA
Lettvin JY, Maturana HR, McCulloch WS, Pitts WH (1959) What the frog's eye tells the frog's brain. PIRE 47:1940–1955
Marr D (1982) Vision. W. H. Freeman & Co, San Francisco
Maturana HR, Varela FJ (1980) Autopoiesis: the organization of the living. In: Maturana HR, Varela FJ (eds) Autopoiesis and cognition: the realization of the living, Dordrecht, Reidel, pp 59–138
McDowell J (1994) The content of perceptual experience, in his mind, value, & reality, Harvard University Press, 1998, Cambridge, MA, pp 341–358
Millikan R (1984) Language, thought and other biological categories. MIT Press, Cambridge, MA
Ryle G (1949) The concept of mind. Penguin, 1963, New York
Tinbergen N (1963) On aims and methods of ethology. Z Tierpsychol 20:410–433
Wittgenstein L (1953) Philosophical investigations, Oxford, Blackwell (trans. of Philosophische Untersuchungen by Anscombe GEM, edited by Anscombe GEM, Rhees R)