SWIF Philosophy of Mind Review
http://lgxserver.uniba.it/lei/mind/swifpmr.htm

Symposium on *Philosophy of Psychology: A Contemporary Introduction*

by José Luis Bermúdez

Edited by Carlo Gabbani

Vol. 5 - No. 3 - 2006
http://lgxserver.uniba.it/lei/mind/swifpmr/0520063.pdf

SWIF Philosophy of Mind Review

http://lgxserver.uniba.it/lei/mind/swifpmr.htm

An on-line publication of the
SWIF Sito Web Italiano per la Filosofia
(Italian Web Site for Philosophy)
http://www.swif.it/

ISSN 1126-4780

General editor:
Luca Malatesti (University of Hull)

Contents

Rise or Fall of the Philosophy of Psychology?

Carlo Gabbani

Department of Philosophy
University of Florence

Over the past years topics in the philosophy of psychology have been increasingly discussed in scientific debates. Several general introductions to this field of study have been published in different languages and a lot of issues coming from this area are present even in the innumerable books on philosophy of mind that have been printed over the last decade. It is also for this reason that a debate on the status, methods and aims of this discipline can't be postponed. As such, the recent general introduction to the philosophy of psychology by José Luis Bermúdez seems to be an ideal starting point. This is especially the case since the panel of contributors on this *Forum* is made up of scholars that are not only distinguished philosophers from various countries, but most of them are also authors of general introductions to the philosophy of psychology (see: Botterill-Carruthers 1999) or to the philosophy of mind (see: Lowe 2000; Paternoster 2002; see also: Id. 2005).

Bermúdez's book represents a very remarkable and up-to-date synthesis of a large number of issues which are relevant in the philosophy of psychology, a discipline characterized by the author (also in order to distinguish it from the philosophy of mind) as *"concerned primarily with the nature and mechanisms of cognition, rather than with the metaphysical and epistemology of the mind"*. At the same time, for Bermúdez, this discipline *"lacks the insulation from scientific research and concerns that more traditional debates in the philosophy of mind possess in virtue of their metaphysical and epistemological dimension"* (p. 15).

The introduction he offers to this field of studies consists in ten chapters and moreover some interesting *'Concluding thoughts'* which Bermúdez offers as the opening-text of this *Forum*.

It would be impossible to provide here a detailed summary of the different topics discussed in the book. But as illustrations, I can mention:

-the original analysis of the connections between perception, knowledge and action and how this bears upon the modularity of our mind. More precisely, a new perspective is offered concerning the relationship between peripheral processing and other non-peripheral types of processing, both at lower or higher levels (especially in chap. 8);

-the discussion of the relationship between thinking and language and a careful presentation of the so-called 'rewiring hypothesis' (especially in chap. 9 and 10).

Moreover, a special attention is justly devoted throughout the whole book to the problem concerning the nature, aim and limits of horizontal *vs.* vertical description/explanation of mental events and intentional behaviour. In the first case (a horizontal explanation) a mental state or event is explained with isomorphic terms, calling only on other events and states of the same kind: from this perspective a public event described within the framework of social sciences, for instance, is explained in terms of antecedent events of states of the same level, described within the same language, or a neural process is explained with reference to the knowledge of other processes at a neuroscientific level, involving the interactions between areas and neurons.

In contrast, with a vertical explanation we search for an answer to our questions at a (supposed) lower, more fundamental level of explanation (and of reality) in comparison to the very level of our original *explanandum*. In this way, for instance, one can explain a person's beliefs in terms of neuro-biological processes and causes present in their brain.

It's easy to understand that the decision regarding this problem, i.e. the coexistence or the clash of these models of analysis, amounts to the adoption of a definite position regarding the nature and effectiveness of categories and statements of commonsense psychology, which is (according to Bermúdez) the highest level of analysis for psychological phenomena. As such, it is also the first level of analysis questioned as one calls on the different types of vertical analysis. In other terms the problem of horizontal *vs*. vertical explanation is, in this domain, first of all the problem of a philosophical evaluation of personal and/or *sub*-personal levels of explanation.

Central to the book is, therefore, the articulation of what are, for Bermúdez, four different, alternative, pictures of the mind, centred also on different ways of thinking about the relationship between different levels of explanation. These are the *representational mind*, the *functional mind*, the *autonomous mind* and the *neurocomputational mind*. Such images are introduced from the first chapters and the dialectic among them is the main thread throughout the book, because each image with its value and limits can be viewed at the same time either as a valuable alternative to other images already in existence, or as a cause of discontent which subsequently generates other pictures.

Bermúdez stresses that each picture seems to work well for a specific, limited aspect or for a particular dimension of our mind. None of these are completely useless, but at the same time none of them are completely persuasive for each and all characters of mind. From this perspective Bermúdez takes stock of the situation with the hope of moving towards a fifth picture, different from every *"monolithic account"*. This new picture (see pp. 318-332 of the book, and Bermúdez's text for this Forum) also represents an attempt *"to explore the possibility of combining some of the insights and analyses offered by the different approaches"* (p. 320) which nevertheless gets over their limits, especially those concerning the problem of the *interface* (as he says) between different explanatory levels (see: § 2.4).

But at the same time the purpose of this picture is not only to be a 'hybrid' or 'irenic' picture, but instead just a new, original 'fifth picture', among the main characteristics of which I would like to underline three points:

1. *"to circumscribe the role of the propositional attitudes, rather than to banish them altogether"* This is done especially by breaking *"the connection between intelligent behaviour and the propositional attitudes"* (p. 322); and our understanding of other people would also be from this perspective largely independent of propositional attitudes.

2. More in detail: *"The massive modularity hypothesis opens up the possibility of more or less direct links between perception and action that are sophisticated enough to be characterized as forms of intentional behaviour, and yet that do not engage the propositional attitude system"* (p. 242);

3. At the same time: *"a natural language medium is required for all types of thinking that have a* metarepresentational *component, that is to say, all types of thinking that involves thinking about thinking"* (p. 328; see also pp. 287-295).

It is exactly this open-conclusion that has represented the main stimulus for our debate which has concentrated especially on the 'interface problem' and on the solution offered to it by Bermúdez's fifth picture of the mind.

- George Botterill's paper discusses the four pictures sketched by Bermúdez and criticizes the idea of focusing the debate on the 'interface problem', since it would provoke *"a* *rush to argumentative engagement on the basis of an inadequate descriptive account"*. This inadequacy would be especially connected to the wrong conviction that we have a natural, obvious insight into the nature of our common psychological abilities, and that this can be understood in terms of a theory called 'folk psychology'. On the contrary, he suggests, there is a hard, interdisciplinary work to be done in order to reach a better understanding of our common-sense psychology, before an interface problem can be properly analysed. Besides, he argues that Bermúdez's fifth picture constitutes more than a 'fifth picture. It's really a third framework (in addition to Fodor's model and the massive modularity hypothesis) concerning our cognitive architecture.

- Jonathan Lowe's contribution is a defence of the 'autonomous mind', a mind that is indeed even more autonomous than the mind sketched by the authors Bermúdez discusses as representative of the autonomous picture (such as J. McDowell and J. Hornsby). Lowe defends a non-cartesian form of psychophysical dualism and proposes an account of voluntary and intentional human action, trying to explain how, from his perspective, mental and neurophysiological causes interrelate with one another. He argues on the one hand that it is impossible to identify the agent's intentional act with any 'blind' neural event (or with a combination of them). On the other hand he proposes a counterfactual-based argument against the psychoneural causal identity, that is an argument which aims to show that a decision *D* cannot be identical to the neural event *N* (with which the physicalist proposes to identify it), as the counterfactual implications of the non-occurrence of these two events would be different.

- Alfredo Paternoster's paper also deals primarily with the interface problem. He believes that this problem represents a very relevant problem (also) at a conceptual level and that it can reveal itself to be more complex than Bermúdez presumes. In particular, if within Bermúdez's fifth picture the role of propositional attitudes is restricted, but not questioned in itself, the interface problem, as a conceptual problem, seems to remain untouched in every case in which folk psychology is relevant. As he says: *"if the interface problem has to be considered as a serious problem, narrowing its scope makes it less pressing, but in no way less difficult"*.

- Karen Shanton focuses on another aspect of Bermúdez's fifth picture, that is the idea that the possession of propositional attitudes depends (in different ways) on the acquisition of a natural language. From this perspective the lack of natural language also seems to imply the lack of any propositional attitude. But Shanton suggests that recent empirical findings seem to cast doubt on this idea: animals and very young human infants seem not to possess natural language at all, but at the same time a lot of scientist seem to suggest that they do have propositional attitudes. Shanton‘s paper discusses whether this kind of evidence can be considered to break the link posed by Bermúdez between language and propositional attitudes, and she raises doubts on the possibility of consistently denying that such categories of non linguistic beings possess true propositional attitudes at all.

I would like to add some brief further remarks and especially two general observations concerning the status and aims of the philosophy of psychology.

More in particular, I would like to suggest a reflection on the possibility that, in order to further its achievement, the philosophy of psychology should on the one hand (a) enlarge its range of interests and field of analysis, and (b) on the other hand defend the conceptual autonomy of its own paradigm.

(a) To amplify the range of interest of the philosophy of psychology probably means asking ourselves why we should be essentially interested only in cognitive mechanisms as objects of reflection. Why, for instance, should we not recognize an analogous space or relevance to the analysis of emotions, or to the philosophical analysis of mental disorders? Or that of self-deception? or brain damage?

Obviously the point should not be to focus on what is not present in a very ample book, instead of on what is present (and moreover these absences are very common in books of this kind). The point is rather to evaluate: (i) whether or not what has been proposed (see: p. 15) as definitive of what the philosophy of psychology should be, is adequate or too restricted; and (ii), more specifically, whether or not the very analysis of cognitive mechanisms can really be secluded from what pertains to the spheres of emotions, mental illness, mood disorders and so on (only as an example of the relevance that an analysis concerning these kinds of dimensions, see Bolton-Hill 1996 especially on mental causation and intentionality).

(b) But even more relevant seems to be the question concerning the autonomy of the philosophy of psychology, that is the possibility of a philosophy of psychology *in the "strong" sense*.

What is intended by the expression "philosophy of psychology in the strong sense", or in the proper sense is an analysis which is not only the most general branch of psychology itself or a taking note of the general interdisciplinary consequences of experimental research concerning cognitive processes (philosophy of psychology *in the "weak" sense*), but which is also a critical, philosophical exercise applied to psychology (even if it can be free from all metaphysical commitments). The aim of this reflection will not be to give precepts, but to conceptually analyse categories, methods, answers, levels of analysis and the cognitive reach and value of empirical evidence in relationship to traditional, philosophical problems. The problem of the status of the philosophy of psychology will then be connected to the wider question of the status and fate of epistemology, and to how the latter relates to the empirical science of nature (more in particular to that of human nature).

In my opinion, this connection raises some puzzling issues about the possibility of a philosophy of psychology kept apart from more general epistemological debates. I say this because I think both that the outcomes of the philosophy of psychology necessarily have consequences for the way we regard general epistemological problems, and, at the same time, I doubt that one can adequately define the conceptual space and the methods of the philosophy of psychology unless one faces up to some general epistemological matters. I have in mind questions (some of which are actually very well analysed by Bermúdez) such as scientific realism, the relationship between observable and inferred entities, the logical relationship between the framework of the *explanandum* and the framework of the *explanans*, the compatibility of a naturalized epistemology with its normative dimensions, and so on. Especially with reference to this latter point it is perhaps just the philosophy of psychology that is able to play a key-role. This can be viewed as the *locus* within which the question concerning the possibility of an epistemology considered as a chapter of empirical psychology, meets the problem of the possibility of an epistemological reflection on empirical psychology itself.

From this perspective, I find it above all important to examine whether the increasing interest in the philosophy of psychology is not too heavily characterized by the presence of research projects and epistemological paradigms that could potentially cause the decline of the possibility of an autonomous philosophical analysis of the empirical search. That is to say, a potential crisis for the very existence of a philosophy of psychology in its strong or proper sense.

As I have said, I see in the philosophy of psychology a very peculiar and close bond between the method and the matter. In particular, I would be inclined to presume that the very possibility of a philosophy of psychology also depends on the fate of the 'personal level of analysis', sketched by Bermúdez in connection with the autonomous picture of the mind (I can't question here his characterization, see on this point Lowe's paper).

We can probably summarize one main aspect of the problem concerning this, saying that for some scholars the 'personal level' of description and explanation reveals itself as being the *locus* in which plenty of errors and theoretical misconceptions concerning the real nature of our mental life are located. We should therefore abandon this level, reconfiguring scientifically our understanding of the conscious experience. From this perspective the impossibility of an ideal intertheoretical reduction is no more than a supplementary proof of the radical falsity of the common-sense image of our psychology. Other scholars, on the contrary, not only maintain that our abilities concerning self-knowledge, mind-reading, prediction and planning of behaviour (i) are essentially connected to the high and personal level of analysis, and (ii) are extremely reliable. But, above all, they say that from a conceptual point-of-view, the personal level of analysis represents the only possible level within which some very relevant phenomena of our conscious experience are really 'on view'.

Now, I would like to suggest that the possibility of certain mental phenomena being irreducible to a subpersonal level of analysis, might not be only an obstacle to solving the 'interface problem' and consequently for the achievement of a philosophy of psychology (as Bermúdez seems to suggest, see: p. 51), but rather a precondition for the possibility of the existence of both (the interface problem and the philosophy of psychology).

To tell the truth, I don't deny the possibility that certain versions of the defence of an 'autonomous mind' could represent a limitation for the development of a philosophy of psychology, but I think this is not something necessarily connected to a fair estimation of the value of the personal level of analysis of the mind. Obviously, to underline the necessity of a personal level and of autonomous dimensions of mind does not amount to the idea that every aspect of our mental activity develops at a personal level, or even only that this is true for all aspects pertaining to our intelligent (both practical or epistemological) behaviour. There are plenty of mental events going on at a subpersonal or at an unconscious level: and they are obviously in a certain way relevant even for the phenomena characterising each individual *personal* manifestation.

Now, the 'interface problem' between personal and subpersonal dimensions, and the articulation of 'enabling conditions' might indeed be considered difficult problems. But to overcome these difficulties dissolving one of the poles of this problematic link, represents, in my opinion, too high a price in order to reach an explanation: because it would be necessary to 'constrain' the *explanandum* beyond what is suitable, that is almost to the point of dissolving it (a similar point seems to have arisen in Paternoster's paper). In other words: I'm not sure that the reflections elaborated up to now on the 'interface' between different levels are completely unsatisfactory (and new and different proposals about this problem are also presented in this Forum). But above all, I fear that the complete denial of a personal, irreducible dimension of mind would represent a gamble to the very possibility of significant explanations of our experience and (at the same time) of an epistemology (I must underline that this complete denial *is not* Bermúdez's choice, because he proposes instead an overall re-evaluation and narrowing of the nature and length of the personal level of analysis; see for instance: pp. 242-243 and 323).

More in detail, I suggest that a serious difficulty both for a genuine philosophy of psychology and for genuine explanations of our conscious experience, would be created by the conjunction of a fully naturalized epistemology with the so called eliminative revisionism (on this theory see also: Botterill-Carruthers 1999, chap. II; Gabbani 2007). We deal here with a (radical) version of the 'neurocomputational mind' (see Bermúdez's book chap. V) according to which the *theory* of mind that common-sense psychology would be based on, is false up to the point that a perfect inter-theoretical reduction from entities of the 'mentalistic' language to that of the neurocomputational framework is impossible, and we should rather simply abandon to a great extent the 'folk' picture of our mental activity with its categories.

In the case of such a 'monolithic' version of this option (that is in the case where science replaces the personal-level account or produce an unfair 'co-evolution' of it), I think we would have a situation in which the philosophy of psychology would result in being no more than the more general chapter of experimental psychology, without a peculiar identity and autonomous problems.

And this explains why Patricia Churchland explicitly says that from her ‘neurophilosophical’ point-of-view she finds fundamental questions of epistemology as: *“the Gettier problem, the nature of sense-data, the nature of incorrigible foundations of knowledge, and the constituents of the corpus of a priori knowledge”* no more than *“old curiosities”* (Churchland 1987, p. 544).

The debates on the a priori or on the Gettier-problem have been more active then ever over the last decades, and there is no need to defend them. But while the beneficial influence that developments in empirical research do have on these problems is an historical fact, I fear that such radical, dismissive, statements can, on the contrary, be considered as representative of a new form of the *“myth of the given”*. I mean the idea that results of experimental inquiry can situate themselves with their own hands in the conceptual space of philosophical problems, doing therefore the work that epistemology has done (or tried to do) in the past. From this perspective (or if epistemology could be only *“experimental epistemology”*, using V. S. Ramachandran’s words), I can’t see a very significant role for a philosophical reflection on psychology.

But just these words by Patricia Churchland do show again how largely the evaluation of results within experimental psychology, presupposes and implies general epistemological considerations.

This means that *paradigms within* the philosophy of psychology are not irrelevant for the *possibility of* a genuine philosophy of psychology, as long as certain decisions on that matter can become at the same time a decision on the methodological possibility of the discipline itself.

And in my opinion good examples of such decisions arise with regard to the epistemological relevance and analysis of the so-called ‘autonomous mind’: the question concerning, for instance, the existence and value of an irreducible personal level of analysis (see also the introduction to this topic in Hornsby 2000); or the analysis of the idea that all mental entities in common-sense psychology can be viewed as theoretical entities individuated only through the attributive description of their causal role provided by the best scientific theories (as in Lewis 1972 use of the ’ramseyfication’ model); or the evaluation of the limits of the descriptive and explanatory adequacy of physicalistic vocabularies in relationship to our experience (see: Crane 2003).

Finally, I have the impression that where the dear old epistemological matters are not considered any more than *“old curiosities”* and where there is no more a personal dimension of analysis (that is also the ‘conceptual space’ of the epistemology), the philosophy of psychology risks revealing itself to be no more than a transitory figurehead in the revolutionary transition from *“the manifest image of man in the world”* (W. Sellars) to the radical re-comprehension of ourselves in neuroscientific terms: as a ladder that is to be abandoned once we have climbed up it (that is to say once that the *“clash”* between the two images has resolved with the winning of the latter).

On the contrary, if the personal level of analysis has a peculiar, irreducible role, the philosophy of psychology has a proper ‘space of reasons’, and, therefore, just it can be entrusted with the job of distinguishing and at the same time identifying the interface between what pertains to different levels and different theories, in the general explanation of our experience. From this perspective also the interface problem or the reflection on the ‘enabling condition’ (as long as it’s accessible to human subjects) does not represent only a big difficulty, but can be viewed also as a vital space for the development of the philosophy of psychology and its interdisciplinary analysis.

The emergence of an hard line of inquiry aiming to constrain the *explanandum* constituted by our conscious experience up to the point of having a radical reconfiguration of it by the supposed *explanans*, has caused, I fear, a radical difficulty in *“saving phenomena”*, concealed by the idea that according to Wittgenstein represents the dangerous charm of every reductionism: that is the idea that a certain phenomenon is in fact no more than another (putative) more basic phenomenon *(“The attraction of certain kinds of explanation is overwhelming. At a given time the attraction of a certain kind of explanation is greater than you can conceive. In particular explanations of the kind "this is really only this”* (1966)).

A certain irreducibility and reliability of the personal levels of analysis, instead, seems to constitute not only a premise for the birth of all relevant psychophysical enquiries, but also a warrant for their actual significance for us, *as subject of experience*. Exactly for this reason, I presume, the personal level of analysis (and the 'autonomous picture' of the mind *as long as* it is necessarily connected to the existence of the former), shouldn't only represent just one perspective on the mind among others, but something more basic and relevant even for the life of the other (legitimate and welcome) pictures of the mind.

This doesn't necessarily entail the infallibility of first-person self-ascriptions, or the independence of these statements from any other piece of knowledge, or their foundational role at an epistemic level, as a self-guaranteed form of knowledge: no *"myth of the given"* here. And this doesn't mean either an impossibility of a certain evolution of the common conception of the mind, due to the development of neuroscientific inquiry. What is essential is, rather, to maintain that knowledge and interpretations concerning subjective mental contents, can't be tackled *simply* by their complete reduction to internal, sub-personal questions of the empirical sciences of nature.

Now, just the role of saving the logical and conceptual autonomy of each phenomenon, and at the same time also of clarifying the relationship among different and irreducible levels and disciplines in order *"to understand how things in the broadest possible sense of term hang together in the broadest possible sense of the term"* (W. Sellars 1963, p. 1), can be considered (as long as it pertains to the mind) as a very relevant goal of the philosophy of psychology.
And also for this reason, I suppose, the philosophy of psychology is here to stay.

References

Bermúdez J. L. 2005. *Philosophy of Psychology: A Contemporary Introduction.* London: Routledge.

Bolton D., Hill J. 1996 . *Mind, Meaning and Mental Disorder. The Nature of Causal Explanation in Psychology and Psychiatry*. (new ed.: 2004) Oxford: Oxford University Press.

Botterill G., Carruthers P. 1999. *The Philosophy of Psychology.* Cambridge and New York: Cambridge University Press (Italian translation: *Filosofia della Psicologia.* Milano: il Saggiatore, 2001).

Churchland P. S. 1987. "Epistemology in the Age of Neuroscience." *Journal of Philosophy*, 84: 544-553.

Crane T. 2003. "Subjective Facts.*"* In H. Lillehammer and G. Rodriguez-Pereyra (eds.), *Real Metaphysics Essays in honour of D. H. Mellor.* London-New York: Routledge, pp. 68-83.

Gabbani C. 2007. "A Critical Analysis of the 'Eliminative' Stance (from an epistemological point of view)." In M. Beaney, C. Penco, and M. Vignolo (eds.), *Mental Processes: representation and inference.* Cambridge: Cambridge Scholar Press.

Hornsby J. 2000. "Personal and Sub-Personal: A Defence of Dennett's Early Distinction." *Philosophical Explorations,* 2: 6-24.

Lewis D. K. 1972. "Psychophysical and Theoretical Identifications." *The Australasian Journal of Philosophy,* 50: pp. 249-258. Reprinted in D. K. Lewis, *Papers in Metaphysics and Epistemology.* Cambridge: Cambridge University Press, 1999, pp. 248-261.

Lowe E. J. 2000. *An Introduction to the Philosophy of Mind*. Cambridge: Cambridge University Press.

Paternoster A. 2002. *Introduzione alla Filosofia della Mente*. Roma-Bari: Laterza.

Paternoster A. 2005. "Filosofia del Linguaggio e Della mente: a Cavallo del Secolo." In T. Burge, *Linguaggio e Mente.* Genova: De Ferrari, pp. 75-127.

Sellars W. 1963. "Philosophy and the Scientific Image of Man.*"* In W. Sellars, *Science, Perception and Reality.* London: Routledge, pp. 1-40 (reprinted: Atascadero: Ridgeview, 1991).

Wittgenstein L. 1966. *Lectures and Conversations on Aesthetics, Psychology, and Religious Belief.* Ed. by C. Barrett, Oxford: Blackwell (Italian translation: *Lezioni e Conversazioni sull'Etica, l'Estetica, la Psicologia e la Credenza Religiosa.* Milano: Adelphi, 1967).

The Philosophy of Psychology: Towards a Fifth Picture?

José Luis Bermúdez

Department of Philosophy
Washington University in St. Louis

My book, *Philosophy of Psychology: A Contemporary Introduction*, approaches the philosophy of psychology through the lens of four dominant pictures of the mind. Each picture incorporates a different set of metaphors and tools for thinking about the mind and how it relates to the brain and to the environment. Each highlights different aspects of the mind and offers a distinct way of responding to what I call the *interface problem*. This is the problem of explaining how commonsense psychological explanation interfaces with the explanations of cognition and mental operations given by scientific psychology and the other cognitive and behavioral sciences.

The *representational picture* is built around the metaphor of the mind as computer, treating cognitive abilities in terms of computational tasks and using the idea of computation as the thread linking together different levels of explanation. According to the *functional picture*, in contrast, the causal dimension of the mind is paramount. Instead of focusing on particular cognitive abilities, the functional picture highlights the causal dimension of individual mental states, using the role/realizer relation to show how what goes on at lower levels of explanation can be causally relevant to the personal-level states of commonsense psychology. While the functional and representational pictures try to tackle the interface problem head on, the pictures of the *autonomous mind* and the *neurocomputational mind* try in their very different ways to undercut its force. The picture of the autonomous mind highlights what it takes to be the uniqueness and irreducibility of personal-level psychology, deriving this uniqueness from the norms of rationality claimed to govern personal-level psychology. The picture of the neurocomputational mind, in contrast, is strongly committed to the metaphor of the mind as brain and accepts that our thinking about the mind must co-evolve with our thinking about the brain in a way that may lead to significant revisions of our commonsense ways of understanding cognition and behavior.

Each picture of the mind emphasizes different aspects of cognition and highlights different paradigms. The neurocomputational picture, for example, stresses what one might think of as low-level cognitive mechanisms. It takes issue with the natural assumption that high-level cognitive achievements must be carried out by complex computational mechanisms. Instead, it emphasizes the explanatory power of surprisingly simple mechanisms performing operations of template-matching and pattern recognition. The plausibility of the neurocomputational view is in large part a function of how convinced one is by neural network models of higher cognitive abilities (and indeed of how representative one takes neural networks to be of neural functioning). The autonomy view, on the other hand, takes as its paradigms of cognition the most sophisticated forms of rational reflection and deliberation. The types of thinking highlighted by the autonomy view are not simply *governed by* norms, but rather *guided by* norms in ways that involve reflecting on the demands imposed by norms of rationality. The representational and functional pictures fall somewhere between the two. One basic idea behind the representational approach is that formal transitions between syntactic entities can track semantic transitions. This is of interest primarily in connection with types of thinking that lend themselves to being codified in formal models such as expected utility theory or deductive logic. Whereas the representational picture sees thinking in primarily logical terms, the functional picture takes a causal view of the dynamics of thought. The paradigm for the functional picture is the interaction of beliefs and desires in the generation of behavior. Representational theorists take the challenge to be explaining how logical transitions can be captured by causal transitions. Functional theorists, in contrast, take causal transitions between mental states as basic and see the challenge as showing how those causal transitions can be used to characterize the mental states featuring in them.

Each picture tries to show that the mind as a whole should be understood on the model of the favored paradigm types of thinking. It is predictable where the difficulties will be found. One might reasonably think, for example, that the neurocomputational approach will have difficulties with the deductive transitions and probabilistic calculations taken as paradigmatic by proponents of the representational mind. It is true that theorists probably underestimate the extent to which logical reasoning is a matter of pattern recognition – after all, one can only apply formal rules if one can identify which formal rule is salient in a particular context, and this is often a matter of seeing what pattern is exemplified by a given inference. But it seems likely that the rule-governed nature of logical reasoning will make it difficult to capture with the resources of the neurocomputational approach. By parity of reasoning one might expect the perceptual and recognitional abilities highlighted by the neurocomputational approach to pose problems for representational theorists. Even though perceptual processes are no doubt governed by rules, these rules seem fundamentally different from the inflexible and formal logical rules that are easily captured and manipulated in the language of thought. It is certainly true that researchers in traditional artificial intelligence (what is sometimes called "good old-fashioned artificial intelligence") have had far more success in modeling formal and semi-formal types of cognition that they have had in developing models of perceptual processing.

Similar difficulties arise with the different emphases and priorities of functional and autonomy theorists. Surely, autonomy theorists will ask, there must be more to theoretical deliberation and practical reasoning than causal interactions between mental states. How can a purely causal story can do justice to our more reflexive and reflective modes of thinking? And of course the same problem arises in the other direction. The rarified approach proposed by autonomy theorists seems to involve too much heavy-duty machinery to provide a plausible account of the myriad of trivial inferences and uncomplicated predictions that make up daily psychological life. How much time do we really spend thinking about "how things ought to be", as opposed to making quick and efficient guesses about "how things are".

It has not gone unnoticed that the general approaches to the mind we have been considering each work best for a limited domain. One obvious response is to try to show that thinking and cognition are really far less varied than they initially appear. So, for example, a neurocomputational theorist might attempt to show that cognition is far less rule-governed and language-dependent than it initially appears to be, while a functional theorist might try to show that the norms governing practical reasoning and deliberation can be understood in causal terms. Another response would be to try to finesse the situation by locating different approaches at different levels of explanation. For example, supporters of the representational approach standardly argue that it is not directly in competition with the neurocomputational approach, because the neurocomputational approach is best viewed as an account pitched at the implementational level. Similarly, autonomy theorists frequently argue that the causal approach adopted by functional theorists is best seen as an account of the subpersonal underpinnings of cognition, rather than of personal-level thought.

It seems unlikely, however, that the strategy of either assimilating the competition or trying to show that there is no real conflict by locating the apparent competition at a different level of explanation will prove completely satisfying. Thinking and cognition are just too complex and variegated. In the light of this it is natural to wonder whether trying to find a single monolithic account of the mind as a whole is really the best strategy. Perhaps it would be more profitable to explore the possibility of combining some of the insights and analyses offered by the different approaches. I will make some very preliminary and programmatic remarks about one possible way of developing such an alternative account. What follows draws upon some of the arguments and claims that have emerged in the main body of *Philosophy of Psychology: A Contemporary Introduction*, but is very much a personal view. The suggestions that follow represent one way of navigating through the complex issues in this area, but it is certainly not the only way and there may well be better ways.

Let me begin by drawing attention to some ideas that emerged in the course of the book. One important theme has been the significance of commonsense psychology. All four pictures of the mind we have been examining take commonsense psychology to play a fundamental role in our understanding of ourselves and others. Commonsense psychology is an explanatory tool that explains and makes sense of behavior by interpreting it as the result of beliefs, desires and other propositional attitudes. A commitment to the explanatory power of folk psychology fits naturally with the view that beliefs, desires, and other propositional attitudes are the "springs of action". The simplest explanation of the explanatory success of commonsense psychological explanations is that they work because they are true, which is to say that they work because they correctly identify the beliefs and desires that really caused the actions in question. And similarly for prediction. One might think, therefore, that whenever we are dealing with behavior that cannot be seen as a direct response to some environmental stimulus we must be dealing with action that is in some sense generated by propositional attitudes. As emerges in Chapter 7, this way of thinking about the springs of action brings with it a particular interpretation of the architecture of cognition – specifically, a sharp distinction between "central" cognitive processes that involve propositional attitudes and "peripheral" cognitive processes that are not defined over propositional attitudes but instead provide inputs to the propositional attitude system. These modular processes have certain characteristics (such as informational encapsulation, domain-specificity, speed, and so on) that make it natural to classify them as subpersonal, in opposition to the personal-level propositional attitude system, which has none of these characteristics.

There are ways of putting pressure on this way of thinking about the architecture of cognition. Chapter 6 considers ways of making sense of the behavior of others that do not involve the attribution of propositional attitudes and hence that do not involve the explanatory framework of commonsense psychology. Much of our understanding of other people rests upon a range of relatively simple mechanisms and heuristics that allow us to identify patterns in other people's behavior and to respond appropriately to the patterns detected. The simplest such patterns are a function of mood and emotional state, while the more complex ones involve social roles and routine social interactions. One interesting feature of these modes of social understanding is that, by downplaying the role of the propositional attitudes in social understanding, they diminish the centrality of the interface problem in our thinking about the mind. These are personal-level modes of social understanding that do not bring with them the complicated theoretical machinery that philosophers of psychology have standardly taken to be required for navigating the social world. They do not require maneuvering oneself into another person's perspective on the world (in the manner proposed by the simulationist approach to social understanding), or bringing to bear a tacitly known theory of cognition and behavior (as suggested by theory-theorists).

Of course, our ways of explaining behavior are not invariably a good guide to how that behavior came about. Optimal foraging theory is a striking example, where a complex theoretical framework is used to explain and predict behavior generated by a set of very basic mechanisms and rules. But the discussion of ways of thinking about the path from perception to action in Chapter 7 suggested that there is a range of ways of generating behavior that are neither reflex or instinctual, nor are mediated by propositional attitudes. The line between perception and cognition may not be as sharply defined as is standardly thought. There are ways of perceiving the world that have direct implications for action. Frequently what we perceive are the possibilities that the environment "affords" for action, so that we can act on how we perceive the world to be, without having to form or exploit beliefs and other propositional attitudes. The perception of affordances cuts across the sharp distinction between, on the one hand, peripheral, domain-specific, and informationally encapsulated modules providing a "neutral" representation of the distal environment and, on the other, central cognitive processes defined over the propositional attitudes.

The discussion of the massive modularity hypothesis in Chapter 7 puts further pressure on the standard distinction between peripheral and central processes. According to the massive modularity hypothesis, there is no such thing as domain-general thinking. All thinking is subserved by domain-specific modules that evolved to deal with specific problems confronted by our hominid or primate ancestors. These so-called Darwinian modules are very different from the modules discussed by Fodor. They are not informationally encapsulated, for example, and their principal function is not to transform sensory input into a format that can serve as input into central processing. They are modular in two senses. First, they are domain-specific – engaging only in response to a limited set of inputs and applying only a limited set of operations to those inputs. Second, the representations they employ are not best viewed in terms of the categories of propositional attitude psychology.

How should we respond to these pressures on the standard distinction between subpersonal modular processing and a personal-level propositional attitude system? One response would be eliminativism about the propositional attitudes, effectively holding that the propositional attitudes should have no role to play in how we think about the genesis of behavior –and hence, *a fortiori*, no role to play in social understanding. Such an approach would mesh well with some ways of developing the neurocomputational approach to the mind – in particular with the views put forward by the Churchlands. On the other hand, however, one might wonder whether eliminativism is too drastic a response. Perhaps it would be better to circumscribe the role of the propositional attitudes, rather than to banish them altogether. The most obvious way of doing this would be to break the connection between intelligent behavior and the propositional attitudes by accepting that there are many ways of behaving in a non-instinctual and non-reflex manner that completely bypass the propositional attitudes. These are forms of behavior that we can explain and understand quickly and efficiently without bringing to bear the machinery of propositional attitude psychology.

Of these two possible responses the balance of the arguments in the main body of the book seems clearly to point to the second, less drastic response. It is hard to imagine that all our talk of propositional attitudes will turn out to have been completely mistaken and that all the work that we take to be done by the propositional attitudes will turn out to be performed by Darwinian modules, mechanisms of template-matching and pattern-recognition, and ways of accommodating oneself to established social routines. It is more plausible to think that the propositional attitudes do have a very real role to play in certain types of thinking and in the genesis of certain types of behavior – particularly where we find the types of norm-guided thinking highlighted by autonomy theorists and the logical thinking emphasized in some of the arguments for the language of thought hypothesis.

One might try to accommodate these various pressures at the level of cognitive architecture by revising the standard distinction between central and peripheral processing in favor of a three-way picture distinguishing two fundamentally different forms of personal-level cognition, in addition to the peripheral modules responsible for processing sensory input. Personal level cognition can involve either the complex processes and mechanisms defined over the propositional attitudes or the much simpler Darwinian modules, heuristics, and mechanisms of template-matching and pattern recognition that we have been discussing. The suggestion here is not that we interpose an additional set of mechanisms between peripheral modules and central cognition, but rather that we think of there being two fundamentally different personal-level routes to action, one engaging the propositional attitudes and the other engaging evolutionarily more primitive mechanisms that are faster and more specialized. The standard distinction between peripheral processing and modular processing can be visualized two-dimensionally, as a core of central processing bounded by an input layer and an output layer of peripheral modules. The current view is best construed in three-dimensional terms, with the propositional attitude system superimposed upon a complex network of pathways leading from peripheral input modules to peripheral output modules. Some of these pathways correspond to Darwinian modules and others to heuristics and social routines. Each pathway leads from input modules to output modules without engaging the propositional attitude system. We might think of each individual pathway as working to solve a particular set of problems in response to a particular type of input. It may be, for example, that one of these pathways corresponds to the so-called cheater detection module, processing inputs of social situations to search for free riders. On the view being suggested, the cheater detection pathway does not work to produce beliefs – it does not feed directly into the propositional system. Rather it has immediate implications for action. The problems it solves are problems of how to behave in particular situations. These are problems, crudely speaking, of whether or not to cooperate, with the question of what is to count as cooperation clearly fixed by the context in which the issue arises. Once the cheater detection module has done its work there is standardly no need for further processes of practical reasoning involving the propositional attitude system – although of course there are different ways of reacting to the presence of a free-rider and there has to be *some* way of deciding between them.

Three significant challenges naturally arise at this point. The first is briefly considered in section 7.4 in the context of Fodor's argument against the massive modularity hypothesis. As Fodor points out (Fodor 2002), there is a lack of fit between the outputs of peripheral modules (what we might think of as Fodorian modules) and inputs to Darwinian modules. The Fodorian modules that collectively comprise the early visual system collaborate to produce a representation of the three-dimensional layout of the distal environment that has only a rudimentary degree of interpretation. The cheater detection module, however, requires highly interpreted inputs. It will only work on representations of social exchanges – and indeed only on those social exchanges that have a cost-benefit dimension. Clearly there needs to be some further processing intervening between the end of peripheral processing and the various pathways that we have been discussing. The first issue, then, is giving an account of this processing and how it fits into the overall architecture of cognition. This is not a topic that has received any attention in the psychological or philosophical literature. We are dealing with processing that effects a form of filtering, working to parse and interpret the deliverances of the modular sensory systems into a format that will engage one or other of the Darwinian modules or other pathways from perception to action. As such it will be a form of domain-general processing. However, as we saw in section 7.4, there is no need to follow Fodor in the claim that it will have to engage what he thinks of as the domain-general propositional attitude system. A proper development of the position being sketched out here will need to offer a substantive account of this type of intermediate domain-general processing. It is very possible that research into artificial neural networks will be illuminating in this area. The filtering tasks that need to be carried out at this level may well turn out to involve the type of detection of patterns and sensitivity to prototypes that artificial neural networks are so good at modeling.

We can view the first challenge as demanding an explanation of how a particular form of selection problem is solved. This is the selection problem of determining which of the various possible perception-action pathways should be engaged in a particular context. But this is not the only selection problem that needs to be solved. I have suggested that processes and mechanisms involving propositional attitudes are superimposed upon the more primitive framework of perception-action pathways. But what determines whether and when these processes and mechanisms are engaged? Again, we are not in a position to make anything more than some very general comments. We can view the propositional attitude *complex* (a better terminology, I think, than the widespread talk of the propositional attitude system) as coming into play to deal with situations that cannot be dealt with by the lower-level perception-action pathways. This would occur most obviously when we are dealing with types of thinking that are not a response to particular demands imposed by the immediate environment – forms of reflection, deliberation, and forward planning that are not stimulus-driven. It is no accident that these are taken as paradigmatic types of thinking by those who see the propositional attitudes as central to cognition. But one might also expect elements of the propositional attitude complex to be engaged in the face of stimuli that do not fall neatly into the domain of one and only one perception-action pathway. It may not be possible to parse certain unfamiliar situations into a format that will serve as input into one or other pathway. In such a situation one might expect that background beliefs will need to be brought into play. Conversely, there may be situations that fall within the domain of more than one perception-action pathway – a situation, for example, that comes within the ambit both of the cheater detection pathway and the danger avoidance pathway. In such circumstances the two pathways may come up with different and incompatible actions. The resources of the propositional attitude complex may be required to resolve the conflict. But how does this take place? How are conflicts between perception-action pathways identified? How are unfamiliar situations "handed over" to the propositional attitude complex? These are all questions that call for considerable further study.

The third challenge in this area is to give a principled account of the significance of natural language in cognition – and in particular of the relation between natural language and the propositional attitudes. This is important if we are properly to evaluate the various arguments for the language of thought hypothesis considered in Chapters 8 and 9. The force of those arguments was that the propositional attitude complex must be explained independently of natural language, because we can only give an account of what it is to learn and understand a natural language in terms (*inter alia*) of beliefs about the means of words – beliefs that cannot themselves be in any sense dependent upon natural language. We considered an alternative to the language of thought hypothesis. This is what I termed the rewiring hypothesis, according to which the architecture of cognition is fundamentally changed by the acquisition of language. Learning a natural language makes available a linguistic medium for thinking that can do much of the work that it is claimed can only be done by the language of thought hypothesis, such as for example explaining the apparent systematicity and productivity of thought. The dialectic between the language of thought hypothesis and the rewiring hypothesis is complex, but we can use the proposals about cognitive architecture made above to get them into focus.

It seems clear that the types of information-processing carried out by Fodorian modules have nothing to do with language mastery, except for those directly implicated in language comprehension and production. And let us assume (as seems plausible) that perception-action pathways of the type we have been discussing are equally independent of language. This allows us to formulate what is at issue between the language of thought and the rewiring hypotheses as follows. The rewiring hypothesis is committed to two claims. The first is that we can explain what is going on in peripheral modular processing and perception-action pathways without needing to postulate a language of thought. Modular processing and perception-action pathways may well involve the processing of information, but not in a manner that requires a language of thought. The arguments we considered in Chapter 9 trying to show that the language of thought is implicated in basic perceptual processing are obviously very much to the point here. The rewiring hypothesis stands or falls with the failure or success of those arguments. The tenability of the rewiring hypothesis depends upon being able to develop plausible models of these types of information-processing in terms of mechanisms of pattern recognition and template-matching – as opposed, for example, to the mechanisms of hypothesis formation and testing favored by proponents of the language of thought hypothesis. It is certainly too early to come to any firm conclusions about where the balance of the arguments lies, but let us grant the rewiring hypothesis that there is a plausible story to be told in this area. The next question that arises is whether we can explain what it is to learn and understand a natural language in terms of the same type of mechanisms as are involved in modular processing and perception-action pathways. Here matters are even less clear than they are with respect to modular processing and perception-action pathways.

Very little is known about how languages are learnt and understood. Proponents of the language of thought hypothesis have an *a priori* argument aiming to show that languages can only be learnt through processes of hypothesis formation and testing that require a language of thought – and, moreover, that understanding the meaning of words needs to be modeled in terms of meaning rules formulated in a language other than the language being understood. Against this proponents of the rewiring hypothesis can muster a range of empirical considerations and theoretical arguments. As we saw in section 9.6 there is a range of models of linguistic understanding that do not appeal to meaning-rules of the type envisaged by Fodor, and fairly strong grounds for thinking that the meaning-rules approach cannot work for at least some central cases. In section 4.3 we looked at interesting evidence that artificial neural networks trained to perform language-learning tasks reproduce certain of the learning effects discovered in young children.

It is worth drawing attention to some of the theoretical possibilities opened up by the rewiring hypothesis. The most striking is the possibility of explaining the phenomenon of language in complete independence of the propositional attitude complex. This would allow us to appeal to language in giving an account of the propositional attitude complex. We might think about the vehicles of propositional attitudes in terms of the rewiring of the brain that occurs when language is acquired – as opposed, for example, to thinking of them in terms of physical realizers of functional roles, or sentences in the language of thought. This would open up the way for a version of what in Chapter 4 we described as the co-evolutionary research paradigm. Our thinking about the vehicles of propositional attitudes would co-evolve with discoveries about the changes that take place in neural structure and neural functioning as language develops. This is as yet fairly uncharted territory. Neuroscientists and empirical psychologists have devoted considerable attention to studying the localization of language in the brain, using evidence from lesions and from imaging studies (Garrett 2002). But this research has tended to be insufficiently fine-grained to help with the problems with which we are concerned. The hypothesis is pitched at the level of individual representations – a level at which the appropriate unit of analysis is the small-scale neural population, rather than the functional area. Moreover, the rewiring hypothesis is more concerned with the representational changes that take place within the brain as whole as a consequence of language acquisition – changes that are hypothesized to occur even in areas that are not dedicated to one or other aspect of language processing.

It certainly seems plausible that the ontogenesis of the human infant involves a process of representational change in which types of mental representation of increasing complexity and sophistication become available – and indeed that a comparable process of representational change occurred in human phylogeny. Models of the process of representational change in human infancy have been offered by a number of authors, including Annette Karmiloff-Smith (1992) and Jean Mandler (1992). According to Karmiloff-Smith the progression towards language acquisition in infancy is marked by a series of representational redescriptions in each of which information becomes more explicit and available to be exploited in a greater number of transitions and transformations. Unsurprisingly, the emergence of language is responsible for the most far-reaching representational redescription. According to Karmiloff-Smith, information becomes fully explicit and available for general use within the cognitive system when it is re-encoded in an essentially linguistic medium. Similar themes occur in a number of models of the evolution of hominid cognition. As we saw briefly in section 9.2, authors such as Merlin Donald and Steven Mithen have suggested that the emergence of language makes possible the integration of different bodies of domain-specific knowledge (Donald 1991, Mitthen 1996).

If such accounts are on the right lines then we have a promising way of approaching the rewiring hypothesis. However, none of the authors mentioned has proposed a detailed account of the possible neural correlates of representational change. Such accounts as exist have emerged from neurobiologists. The selectionist approach, pioneered by Changeux (1985) and developed by Edelman (1989), postulates a "Darwinian" process whereby an original multiplicity of representational units (groups of synapses for Changeux, neural circuits for Edelman) is selectively pruned, in response to either/both sensory input and intrinsic factors. Another possibility in this area is that representational change is subserved by a process of parcellation (Ebbesson 1984), whereby selective loss of synapses and dendrites leads to increasing differentiation of the brain into separate processing streams. A proper development of the rewiring hypothesis will very likely require building bridges between the neurobiology of representation and more high-level ways of thinking about the nature of representation and the role of representations in cognition.

It is likely, moreover, that a proper working out of the rewiring hypothesis will involve taking seriously the idea that certain types of thinking are actually carried out in a natural language medium. We saw in section 9.1 that there are considerable difficulties with the idea (what I termed the inner speech hypothesis) that all thinking involves the manipulation of natural language sentences. Nonetheless, as emerged in section 9.2, there are certain types of thinking that arguably require a natural language vehicle. Andy Clark has suggested that natural language is the medium for what he calls *second-order cognitive dynamics* – namely, types of thinking that involve explicitly reflecting on one's own cognitive practices, as when one evaluates the reasoning by which one arrived at a particular conclusion, or explores whether a hypothesis is well supported by the available evidence. I myself have extended this suggestion to argue that a natural language medium is required for all types of thinking that have a *metarepresentational* component – that is to say, all types of thinking that involve thinking about thinking (Bermúdez 2003). Metarepresentational thinking includes what Clark calls second-order cognitive dynamics but extends beyond it to include, for example, thinking that involves ascribing mental states to others (which involves thinking about a thought as the content of another's mental state); that involves conceptions of necessity/possibility and tense (since such notions are best viewed as operators applying to thoughts); and indeed to all types of thinking that involve logic (since logical thought involves reflecting upon the structure and truth-value of thoughts).

The basic argument for the dependence of metarepresentational thinking upon language is that it requires the target thoughts to have vehicles that will allow them to be taken as the objects of thought. Since the paradigm cases of metarepresentational thinking are instances of conscious thinking, these vehicles must be available to conscious thinking. They must, moreover, be vehicles that make the structure of the target thoughts available. This is clearly required, for example, if one is to reflect upon the inferential relations between thoughts.[1] Natural language sentences appear to be the only candidates that satisfy both requirements. Other candidates satisfy one requirement, but not the other. Imagistic representations, for example, are consciously accessible, but do not make the structure of a thought available. Formulae in the language of thought, conversely, make structure available, but are not consciously accessible.

This suggestion about the nature of metarepresentational thinking gives us a further perspective on the project of trying to explain the propositional attitude complex in terms of language. It allows us to see the proposed explanation as having two parts, one focusing on first-order propositional attitudes (those propositional attitudes directed at the world, rather than at one's own thoughts or those of other people). It is to these that the rewiring hypothesis primarily applies. The explanatory task here is to understand how the acquisition of language changes the neural circuitry in a manner that creates potential vehicles for propositional attitudes. The second part of the explanation, in contrast, focuses on second-order propositional attitudes (those involved in metarepresentational thinking). What we are interested in here is showing how these types of thinking involve the explicit manipulation of natural language sentences. In particular, we need to understand the process of manipulating natural language sentences in a way that avoids the problems confronted by the inner speech hypothesis.

Of course, in sketching out the principal claims of this fifth picture of the mind I have concentrated on the benefits rather than the costs. And there are a number of significant outstanding problems that will need to be resolved before the prospects can be viewed in as rosy a light as I

[1] Strictly speaking, this requirement holds only for those inferences that exploit the internal structure of a thought – the type of inferences that are the subject of the predicate calculus. Inferences of the type codified in the propositional calculus depend solely upon the truth-values of the relevant thoughts.

have presented them. Some of these we have already discussed – such as the problem of giving a non-metaphorical account of what it is to manipulate a natural language sentence in thought, and the problem of turning the rewiring hypothesis into a substantive theory of the vehicles of first-order propositional attitudes. There is a further problem directly related to an important strand in the arguments for and against the language of thought hypothesis discussed in Chapters 8 and 9. The proposal here is effectively to understand "central" cognition in terms of natural language. Any such proposal has to answer the obvious challenge of explaining what is going on in apparent cases of "central" cognition in creatures that do not possess a natural language. It is well known that cognitive ethologists, developmental psychologists, and cognitive archeologists use the language of propositional attitude psychology to characterize the cognitive abilities of non-linguistic and infra-linguistic creatures and to explain their behavior in both natural and experimental settings. How should we deal with talk of animal beliefs, or infant knowledge? Here we seem to have examples of propositional attitudes that cannot be understood in terms of language and hence that do not fit the proposed model.

One obvious way of dealing with this potential difficulty would be through the minimalist strategy of refusing to take at face value the explanatory practices of cognitive ethology, developmental psychology, and cognitive archeology. Talk of animals having beliefs about conspecifics or infants possessing bodies of knowledge about objects and how they behave should be taken as shorthand for a more complex explanation in terms of the simpler forms of central cognition that we have been discussing. When developmental psychologists analyze experiments using the dishabituation paradigm by attributing to 5-month old infants "knowledge" of the principle that objects move on single connected paths through space-time this should be understood as saying that infants are capable of detecting certain patterns in the behavior of material objects and being surprised by material objects behaving in ways that do not conform to those patterns. Similarly, when ethologists claim that certain species of shore birds set out to "deceive" potential predators by "pretending" to be injured, this should be taken as shorthand for a more complex description of their behavior that can ultimately be understood in terms of innate releasing mechanisms or other, more sophisticated perception-action pathways. Some authors have argued that this type of approach is fundamentally mistaken, on the grounds that we have no better perspective than our actual scientific practices for determining the legitimacy of propositional attitude ascriptions (Kornblith 2002). This may be too extreme, but there is some plausibility in the view that, although one might argue about individual cases, the practice of appealing to propositional attitudes in making sense of the behavior of non-linguistic creatures is too well-entrenched to be dispensed with completely.

Nonetheless, rejection of the minimalist strategy would not leave the defender of the language-based approach to explaining the propositional attitudes entirely without resources. One possible approach would be to exploit the distinction between different types of content that is gaining increasing acceptance. A number of philosophers of mind distinguish between the *conceptual content* characteristic of beliefs and other propositional attitudes, and various types of *nonconceptual content* (see the papers in Gunther 2002). Nonconceptual contents share certain fundamental characteristics with propositional attitude contents. In particular, they can be linguistically expressed by means of "that"– clauses and have a degree of structure that marks them off from perceptual and other imagistic states. What makes them *non*conceptual is that they lack certain fundamental features of propositional attitude contents (with the guiding assumption here being that propositional attitude contents are typically composed of concepts). Most authors who appeal to nonconceptual contents hold that they lack the generativity and productivity generally taken to be characteristic of propositional attitude contents. Since one might well think that generativity and productivity are closely connected with domain-generality, and given that that there is some plausibility (as we saw in section 9.2) in the view that non-linguistic cognition lacks domain-generality, it may well be that we need to characterize the content of the propositional attitudes of non-linguistic creatures in nonconceptual terms. It is natural to combine this with the further thought that we should reserve our propositional attitude vocabulary for states with conceptual content and instead talk of proto-beliefs and proto-desires at the non-linguistic level. Of course, applying the conceptual/nonconceptual distinction in this way would still leave us with the substantive task of making sense of proto-beliefs and proto-desires, but it would allow us to retain the project of explaining the propositional attitude system in terms of language. Nor, one might think, would this be arbitrary or *ad hoc*. The manifest differences between linguistic and non-linguistic cognition make it implausible to think that there is a single category of propositional attitudes that spans both the linguistic and non-linguistic domains.

The possibility is opening up of a picture of the mind completely different from those we have been considering. In place of the standard distinction between input/output modules and a central propositional attitude system, this new picture sees "central" processing in terms of a language-based propositional attitude complex superimposed upon an intricate network of perception-action pathways. The transitions from and to the modular systems on the periphery are effected by systems of domain-general processing that filter the products of modular processing and engage the appropriate perception-action pathways – or, indeed, the propositional attitude system. These filtering systems may well turn out to involve pattern recognition and template-matching of the sort carried out by artificial neural networks. Within the propositional attitude complex we can distinguish two fundamentally different types of cognition. One type of cognition involves first-order, world-directed propositional attitudes and is to be understood at the neural level indirectly in terms of language – that is, in terms of the rewiring that takes place as a function of language acquisition. The second type of cognition involves second-order propositional attitudes, which involve either thinking about thoughts directly, or thinking about the world in a way that requires thinking about thoughts. These are to be understood directly in terms of language, on the assumption that we think about thoughts through thinking about the sentences that express them.

If this picture is viable, then it may well be that we are much closer to understanding the mind than we imagine – or, at least, that we are much closer to having the tools to understand the mind than we imagine. Following on from Marr's pioneering analysis of the early visual system, we have a number of powerful models of modular processing, many of which involve the rapidly expanding resources of computational neuroscience (Churchland and Sejnowski 1993, Eliasmith and Anderson 2002). We also have, in the language of thought hypothesis, an alternative, but nonetheless powerful, theoretical tool for thinking about modular cognition (although, as we have seen, the suggestion that modular processing is a matter of hypothesis formation and testing is far from uncontroversial). It is, moreover, to modular processing that most of the techniques we currently have for studying the brain have been directed. We are moving towards an understanding of the large-scale functional architecture for various types of modular processing, and single-neuron studies have given us some understanding of what is going on at the level of individual neurons. It is true that, once we move beyond modular processing, techniques for directly studying the brain become less relevant. But the rapidly expanding field of research into artificial neural networks offers great promise for understanding the processing required to interpret and filter the products of peripheral modules. Artificial neural networks may also help us to understand what is going on in the various perception-action pathways that we have been considering. As we move "upwards" to the propositional attitude system, the proposal to understand propositional attitudes through the lens of language allows us to apply our understanding of language and language acquisition to try to make sense of the mechanisms of cognition. The benefits are clearest in the case of second-order propositional attitudes, since the proposal is to understand these directly in linguistic terms. It is true that we have as yet very little understanding of how to think through the general implications of language acquisition for neural circuits not specialized for language. Yet the rewiring hypothesis at least offers a way of bringing together what we know (and are continuing to discover) about language in linguistics, philosophy, and the various branches of scientific psychology and using it to inform the study of neural circuits and neural change in neurobiology.

Whatever the fate of the potential approach sketched out in the last few paragraphs, it seems clear that the future of the study of the mind/brain is interdisciplinary. The philosophy of psychology is not just a branch of philosophy that takes psychology and the behavioral and cognitive sciences as its object. It is itself an essential part of the interdisciplinary endeavor of trying to make sense a highly complex phenomenon that can be studied from a vast range of perspectives. As with all multi- and interdisciplinary endeavors, there is an urgent need for a framework that fits together the different perspectives and levels of explanation. It is here that we find the distinctive contribution of the philosophy of psychology – tracing key concepts through different levels of explanation and trying to develop and think through pictures of the mind that tie together the conclusions and techniques of radically different explanatory projects. These are exciting times and, to borrow the words of a well-known philosopher, it is good to know that we are unlikely to run out of work.

References

Bermúdez, J. L. 2003. *Thinking Without Words.* New York: Oxford University Press.

Changeux, J. P. 1985. *Neuronal Man: The Biology of Mind.* Oxford: Oxford University Press.

Churchland, P.S., Sejnowski, T. J. 1992. *The Computational Brain.* Cambridge: MIT Press/Bradford Books.

Donald, M. 1991. *Origins of the Modern Mind.* Cambridge: Harvard University Press.

Ebbesson, S. O. E. 1984. "Evolution and Ontogeny of Neural Circuits." *Behavioral and Brain Science,* 7: 321-331.

Edelman, G. 1989. *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.

Eliasmith, C. and C. H. Anderson 2003. *Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems.* Cambridge: MIT Press.

Fodor, J. 2000. *The Mind Doesn't Work That Way*. Cambridge: MIT Press.

Garrett, M. F. 2003. "Language and Brain." In L. Nadel (Ed.), *Encyclopedia of Cognitive Science*. London: Nature Publishing Group.

Gunther, Y. (ed.) 2003. *Essays on Nonconceptual Content.* Cambridge MIT Press.

Karmiloff-Smith A. 1992. *Beyond Modularity: A Developmental Perspective on Cognitive Science*, Cambridge: MIT Press.

Mandler, J. M. 1992. "How to Build a Baby: II. Conceptual Primitives." *Psychological Review*, 99-4: 587-604.

Mithen, S. 1996. *The Prehistory of the Mind.* London: Thames and Hudson.

Interface and Cognitive Architecture: Do We Understand Commonsense Psychology Well Enough to Tackle the Interface Problem?

George Botterill

Department of Philosophy
University of Sheffield

José Bermúdez coins a seductively memorable phrase — 'the interface problem' — for the issue around which he organizes much of his discussion. This is the problem of how folk (or commonsense) psychological explanation connects (or 'interfaces') with modern scientific explanations of cognitive processes. He describes this as 'one of the key problems in the philosophy of psychology' (p.35). While it is beyond dispute that he has identified an issue which has preoccupied philosophers, there are reasons why it should not be accorded this sort of centrality. The reason I wish to emphasise is that the interface problem, described in these terms, is all too liable to encourage a dangerous tendency in philosophy: a rush to argumentative engagement on the basis of an inadequate descriptive account.

The particular form in which this philosophical weakness is manifested in the case of the interface problem is a presupposition that we already understand well enough, or can sketch out adequately from an armchair position, what commonsense psychology is. This does not seem too outrageous: philosophers are folk, so they should have a grasp of what folk psychology is. What this will come out as, in the practice of the great majority of philosophers (and I am not claiming I can do any better), is belief-desire psychology, with pre-eminence being given to explanation of an agent's actions in terms of her reasons. Nobody is going to deny this form of explanation is an important part of commonsense psychology. But why should we suppose that giving an account of commonsense psychology is a task to which the unaided resources of philosophy should prove equal? While philosophy has something to contribute, giving anything like an adequate account of our ordinary pre-scientific understanding in this domain should be seen as a serious and demanding interdisciplinary undertaking. There is a certain irony here. For any reader is going to be struck by the interdisciplinary spirit of Bermúdez's *Philosophy of Psychology*. It is a work which breathes out interdisciplinary commitment. But in relation to this central concern of the interface Bermúdez falls back on philosophy, and I fear it lets him down. The philosophical failing is a shared, but too shallow, description of commonsense psychology. We will need to trace in a bit more detail how this failing affects the treatment of the interface problem.

Having explored the perspectives provided by four pictures of the mind in the main body of his book, Bermúdez concludes by proposing something he counts as a fifth view. According to this proposal we should think of the mind as a 'language-based propositional attitude complex superimposed upon an intricate network of perception action pathways' (p.331). Since Bermúdez seems to allow that each of the four other pictures can offer a satisfying account either in some domains of cognition or for some aspects of cognition, one might imagine that a worryingly 'Tychonic' proposal is being advanced. 'Tychonic' here is intended as a reminder of an episode from the history of science. Tycho Brahe was the last great pre-telescopic astronomer. His response to the Copernican theory was to accept that other planets orbit the sun, while clinging to geocentrism by placing the sun in orbit around the earth. What may have appeared at the time to be the best of both world-systems is in retrospect a clumsy compromise. Better to let rival research programmes contend on their own terms, rather than patch them together. But Bermúdez's proposal does seem to have its own distinctive claims and commitments. The immediate difficulty is working out how it is supposed to relate to those other four approaches.

It is easy enough to cite distinctive contributions in the style of each of Bermúdez's four pictures. Yet they offer only limited assistance with a general review of assumptions about cognitive architecture, of how many different levels of cognitive processing need to be recognised, and what can be expected in the way of uniformity at any one level. This is clearly because the four pictures are defined in relation to 'the interface problem'. That makes them primarily concerned with the relation between folk or commonsense psychology and how the mind actually operates. While this has been a topic of interest to philosophers, concentrating on that as the central issue may not be the best way of establishing a general paradigm for the cognitive sciences.

There are definitely problems concerning both the independent tenability and the heuristic scope of these four positions. We must pause to enumerate them.

1. The *autonomous mind* picture takes personal-level psychology to be a *sui generis* scheme of interpretation and/or prediction, couched in terms of propositional attitudes and guided by norms of rationality, not answerable or reducible to psychological or neurophysiological accounts of processing at sub-personal levels.

2. The *functional picture* individuates the states of commonsense psychology in terms of their causal roles and interactions, thus leaving every hope that suitable realising states may be found at a sub-personal level.

3. The *representational picture* emphasises the contentful character of psychological states and proposes that transitions between these states are computational.

4. Fourthly, the *neurocomputational picture* is the sole bottom-up methodology in the list, using connectionist techniques and neural networks to make progress with modelling cognitive capacities that we cannot programme directly (e.g., various kinds of pattern recognition and template matching). Bermúdez sometimes calls this the 'coevolutionary view'. It is not obvious why that is an appropriate label. Probably the idea is that it is an approach which might be used to correct and enhance commonsense psychology. In fact its most vocal philosophical advocates have been outright eliminativists about folk psychology.

So much for the manifestos. What do the parties really have to offer? In the first place I cannot understand why we should contrast the functional picture with the representational picture. For not only need there be no contrast or tension between them: it is hard to see how they can avoid being combined, at least in some form. True, it is possible to distinguish functionalist and representationalist theses. But functionalism has a major problem unless it is supplemented by a representationalist view of psychological states. In taking psychological states to be type-differentiated by their roles in cognition, functionalism can only expose roles at the level of the general kind of psychological state: the differences between beliefs, desires, hopes, fears, and the like. But contentful psychological states come in as many different individual manifestations as their content permits, and their interaction with the rest of cognition must be sensitive to variations in that content. Functionalism, therefore, needs to be representationalist. For how else will it be able to accommodate the content-sensitive variation in cognitive interaction within the general functional role of some psychological category (most saliently, belief)?

There is a real contrast with the picture of the autonomous mind. In the course of the book

Bermúdez does not quite commit to an explicit endorsement of this view as the best account of what is going on at the level of the propositional attitudes. But he does not make much attempt to challenge the autonomous mind picture, and I think it is important to do so. There are several reasons for this. One is that a satisfactory account of commonsense psychology requires an understanding of its explanatory structure. This has to be a form of causal explanation, though we need to rejuvenate our model of causal explanation to see this more clearly. (I can only issue a brief suggestion here about the significance of contrastive explanation.) Another reason why I would want to challenge the 'autonomous mind' view is that, as I will go on to argue, commonsense psychology should not be regarded as independent of a more general form of mindreading cognition.

Quite apart from that issue, however, the 'autonomous mind' position simply does not aspire to provide anything like a general cognitive methodology. It does offer heuristics of commonsense psychological interpretation and prediction. But since these apply exclusively to the rational connections between beliefs, desires and actions, it has nothing to offer in the way of a positive heuristic at the sub-personal level. Finally, the neurocomputational approach certainly does have a well developed bunch of heuristics — actually, its major strength. But without any commitment to the level at which it applies, unless neurocomputationalism is pushed in the direction of eliminativism, it is not clear that it can really qualify as a general view of the mind. (This is to cut some big issues about the scope of connectionism brutally short. For we can hardly preclude the possibility that a bottom-up methodology will succeed in a progressive explanatory annexation of higher levels of cognition.)

So perhaps it is a mistake to regard the interface problem as of such central importance in the philosophy of psychology that we should locate general models of cognition in relation to that problem? Actually the interface issue is related in a rather complicated way to the model of cognition Bermúdez outlines in his summary. It is an exemplification and domain of application of that model, but also something more than that: the domain in which the advantages of this model are presented. The model itself, however, is really not so much a fifth picture, as *a third architectural framework*. It constitutes an alternative, not primarily to ways of dealing with the interface problem, but rather to two major and rival views of cognitive architecture. One of these views is Fodor's (Fodor, 1983, 2000), according to which input and output systems are modular, but central cognition is something else (something domain-general, non-modular, and implementationally mysterious). The other chief view of the cognitive system is the Massive Modularity Hypothesis, according to which the mind is modular (in a way) all the way through.

Placed in this context, one can see Bermúdez's proposal more clearly, without the dust thrown up by philosophical debates about the interface. What it amounts to is this: the human mind is capable of handling many cognitive tasks by means of processing which runs straight through from input modules to 'Darwinian' modules to output modules (the perception-action pathways). But it can also send problems which require further and, in particular, domain-general attention into a propositional attitude processing complex, which uses the format of natural language and which is the part of the mind of which we are consciously and reflectively aware.

Now this really is rather Tychonic, because we still seem to have something very similar to Fodorian central cognition superimposed on a modular mind which contains both peripheral input/output modules and other, Darwinian modules. But it would be unfair to speak of Bermúdez's proposal as if it were a fudge. A case can be made out for taking it to be a plausible synthesis. For, on the one hand, Bermúdez can and does take the propositional attitude complex to earn its cognitive keep by enabling us to engage in flexible and inventive feats of problem-solving. In this way his position seems to avoid running into 'Fodor's Problem'. (To use Peter Carruthers' title for the problem of explaining how entirely modular minds could be flexible and creative, in the way that human cognition seems to be.)

On the other hand, since Bermúdez is quite prepared to allow that there are modules which are not merely processing inputs from sensory transducers, he can also accommodate all the evidence and arguments in favour of the modularity of mind. For the great bulk of this evidence, drawn from dissociations, developmental trajectories, and even from the considerations which Cosmides and Tooby appeal to in favour of modular mechanisms shaped by evolution, leads only to *existential* conclusions: that there must be this or that specific module and, cumulatively, that there must be a lot of modules, sitting beyond input modules in processing streams. The only argument for Massive Modularity that Bermúdez cannot endorse is the argument from computational tractability: i.e., that cognitive processing, in order to occur in the natural world at all, must be tractable in terms of computational complexity, and that the only way in which it can be computationally tractable is through occurring within modular systems.

Apart from that specific argument in favour of Massive Modularity, Bermúdez can accept the rest of the pro-modularist case. Doesn't this make his model both plausible and attractive? There are some problems which his position faces, and which he both acknowledges and makes an attempt to engage with in his summary and concluding chapter. I will touch on these, particularly an enforced revision in the scope of application of psychological explanations, in the concluding section. But to assess the case for Bermúdez's view we need to focus more closely on the interface problem by looking at the issues discussed in chapter 7 of *Philosophy of Psychology*. For it is largely through engagement with this problem that Bermúdez's view emerges. Once we have seen the shortcomings of this approach in the domain of mindreading we will be better able to see why the overall model may be less attractive than Bermúdez makes it appear.

A salient feature of Bermúdez's treatment is that his strategy is directed towards reducing the significance of the interface problem, rather than pointing in the direction of a resolution. In general what he does in this chapter on 'The Scope of Commonsense Psychology' is to assume an account of ordinary psychological explanation and prediction as involving a scheme in which propositional attitudes are linked to situations and behaviour. He then points out that explicit and worked out exercises in commonsense psychology are cognitively demanding and likely to be rather rare. He questions whether it is reasonable to suppose that some counterpart or sub-conscious version of commonsense psychology could be operating at an implicit level, using tacit theoretical knowledge. Then he goes on to suggest that we do not need to suppose that commonsense psychology is responsible for our success in coping with social cognition as well as we do, because there are a number of simpler and more direct psychological mechanisms and heuristics which guide us in our interactions in social environments. This general position may appear plausible, and in particular the role assigned to commonsense psychology may seem to accord with intuition.

But it is seriously misleading. In my opinion the defect of this approach is that it takes commonsense psychology to be an independent branch of cognition, while describing it in terms which are really more appropriate to a *practice*. Moreover, such a practice would be better regarded as the socially communicable outcrop of a more basic mindreading capacity. If this is so, then it is important to appreciate where things go wrong.

A first questionable step is made in selecting the label for what we are to examine. Professor Bermúdez must have pondered long and hard over writing of the person-level side of the interface as 'commonsense psychology'. He was faced with a wide choice of terms for designating what we have in this domain: e.g., folk psychology, mindreading, social intelligence, theory of mind (or ToM). It is awkward to make a single selection from the list, because any choice is going to be accompanied by a somewhat contentious theoretical loading. Thus, it seems perfectly sensible to investigate what sort of mindreading abilities other primate species have. But they obviously do not participate in folk or commonsense psychology. 'Folk psychology' has been a favoured term among philosophers, strongly motivated by the functionalist idea that a theoretical structure implicitly defines our psychological concepts. It is hoped we can thereby solve the problem of what gives terms in our ordinary psychological vocabulary their meanings. Hence a philosophical attraction to the theory-theory view (which I freely confess to sharing).

The idea of a folk theory in this context carries some potential dangers, as it can easily lead to interdisciplinary confusions. In talking of 'theory of mind' psychologists may just be intending to indicate a particular domain, without commitment to any one account of how cognition operates in that domain. Developmental psychologists may well be attracted by the idea that cognition in this domain is founded on a theory, or perhaps a succession of related theories. For one thing, that might make acquisition of adult competence easier. Notice, however, that it is by no means obvious how the theory or theories postulated by psychologists in the domain of theory of mind are related to the theory underlying folk psychology which the philosophers want to postulate to make sense of our verbalisable psychological concepts of belief, desire, and the like.

Furthermore, in talking of a folk theory of the mind one might be talking about all sorts of ideas people come to formulate about how minds are to be understood, what Stephen Laurence has rather nicely described (in discussion, AHRC *Culture and the Mind* research project workshop) as 'ideology of the mind'. The point is that it would not be particularly surprising if there were significant differences in such official folk views about how minds are to be understood and the ways in which we really understand minds: the folk, including philosophical folk, may not be the best authorities on their own cognition. This point has serious implications for some of the strategies philosophers have suggested for articulating folk psychology — such as listing all generally accepted platitudes.

In spite of the central role Bermúdez gives commonsense psychology in the philosophy of mind, and in spite of the fact chapter 3 is even entitled 'The Nature of Commonsense Psychology', what commonsense psychology actually is never gets laid out in clear and illuminating terms. This is not really Bermúdez's failing. It comes with adopting the interface problem as an issue, since that comes with a ready-made, but sketchy, characterisation of commonsense psychology as a matter of explaining and predicting behaviour in terms of beliefs and desires. The regrettable truth is that philosophers have made a poor job of describing commonsense or folk psychology. This creates trouble in a number of places, most notably in the discussion of the scope of commonsense psychology in chapter 7.

Early in that chapter (pp.175-6) Bermúdez points to an equivocation in the use of such terms as 'commonsense psychology', 'folk psychology', and 'theory of mind'. He then goes on to distinguish between 'broad' and 'narrow' construals of the domain of commonsense psychology. That would seem to be the main issue in the chapter, an issue which Bermúdez characterizes in terms of 'dominance': Is the employment of commonsense psychology our dominant way of solving problems of social cognition? However, let us consider the alleged equivocation first. The suggestion is that if one adopted a permissive usage in talking of commonsense psychology (or folk psychology) one might just mean whatever capacities and skills we have in virtue of which we can make sense of each other and control our social interactions. Bermúdez remarks: 'In this rather weak sense, it is trivially true to say that all our social interactions are governed by commonsense psychology.' (p.175) He thinks that, having characterised commonsense psychology as whatever enables us to understand one another and cope with social behaviour, there will be no interesting question to raise about the scope of commonsense psychology, and it will surely have to be a large and diverse collection of cognitive capacities.

This looks to be a completely innocent move, a sensible preamble to sorting out what the significant issue is, without prejudging it. Surely that was the intention. And isn't this sort of clarification good philosophical practice? It is philosophical practice. But it isn't always good. The problem is that a conceptual clarification can just shut out something of significant importance. In this instance dealing with the alleged equivocation is a decisive step. For it presents a picture of commonsense psychology, largely a language-based activity, as something distinct from a bunch of lower-level psychological capacities.

What, one might ask, is wrong with that as a clarification of what we are to take commonsense psychology to be? Of what our framework of understanding of the mind is in respect to which the interface problem arises? Bermúdez is going to go on to argue that we do not rely upon commonsense psychology to control all our social interactions, that it is not really deployed all that frequently. Instead he is going to propose certain other psychological mechanisms and heuristics which assist us in navigating the social environment. In this way he appears to be illustrating his general architectural theme. But the terms of the discussion serve to conceal, both from Bermúdez and from his readers, much of what is important about human mindreading.

The crucial feature of the division of commonsense psychology from other psychological processes which Bermúdez suggests are engaged in the domain of social cognition (emotional perception-reaction, the tit-for-tat strategy, frames and routines, pp.199-205; see also p.247: 'a core of propositional attitude psychology, surrounded by a more extensive periphery of heuristics, template-matching mechanisms, scripts, routines, and so forth') is that these other processes are not really forms of *mindreading* at all. But suppose there are some really fundamental mindreading capacities and that some of these inform and guide our ordinary psychological thinking? This is more than a possibility. One thing we know, thanks to the efforts of developmental psychologists, is that ordinary psychology is founded upon an informational understanding of the mind. What is more this informational understanding has a developmental history which predates anything that would naturally be described as 'commonsense psychology'.

We need to recount some of the developmental findings in the domain to see why this is important in the present context. The main experimental paradigm has been provided by the well known false belief task, designed to test whether children are capable of predicting the behaviour of others through attribution of beliefs they do not hold themselves. Many replications of this test established a clear developmental pattern with a step-change normally coming at about four years of age. This has been interpreted as indicating acquisition of the concept of belief required for folk psychology: a metarepresentational understanding. According to one theory what is going on is that a fully metarepresentational understanding of minds supplants an earlier representational understanding in which a child only has a concept of 'prelief', not yet the concept of belief (see Perner 1991). However, there was a suspicion that the apparently clear-cut developmental watershed suggested by the false belief task might be an artefact of the verbal format of the task (see Clements and Perner 1994, 2001, for attempts to detect implicit attribution of false belief). A recent experiment appears to lower the estimate of the age at which children are capable of 'predicting' on the basis of false belief attribution quite dramatically, to fifteen months (Onishi and Baillargeon 2005). This result is still somewhat controversial. For one thing, it depends upon whether looking-time can safely be interpreted as an indicator of pre-verbal 'predictions' (or expectations).

This particular result is too recent to have influenced Bermúdez's position, and there is still room for debate over what it establishes about children's representational or metarepresentational understanding of mind. However, we can even shelve any complications concerning metarepresentation by taking a more inclusive view of mindreading. Children who pass the false belief task are usually able to justify their answers in a way that shows they are already at an early stage of participation in commonsense psychology. But the point that should not be overlooked is that long before then they have been treating other people as having minds, in the sense of having information which can direct behaviour. This is exhibited in capacities known to be impaired in autism, as indicated by lack of protodeclarative communication, gaze following, and shared attention (Baron-Cohen et al., 1996; Baron-Cohen and Swettenham 1996; Baron-Cohen et al., 2000). Protodeclarative pointing and utterances (as contrasted with protoimperative demands, which might be issued just because they are found to be rewarded) are particularly impressive, because the intention of such behaviour is precisely to affect another mind. Commonsense psychology is founded upon these basic capacities for representing information in other minds, and we know that impairment in these capacities leads to persistent problems in social cognition for the autistic.

Taking mindreading capacities which precede and underpin commonsense psychology into account disturbs the application of Bermúdez's general scheme. He is arguing that commonsense psychology is part of the propositional attitude complex (and also a reflection of that complex) and as such gets invoked comparatively rarely in dealing with problems of social cognition:

> '...the paradigms of folk psychological explanations given by theory-theorists tend to be complicated inferences ... These are striking cognitive achievements, but it seems odd to take them as paradigms of interpersonal cognition.' (p.194)

The claim that we are not aware of thinking explicitly in these terms for many minutes each day is plausible enough. It is, however, definitely an empirical matter and we should be wary of settling it by armchair introspection. It would be better to get data from empirical sampling techniques on the question:

> How often and for how long are participants engaged in explicit thinking about behaviour in terms of the concepts of commonsense psychology?

Lacking at present the ingenuity to see how to design anything like a reliable trial of this question, suppose we accept that we do not normally spend much time over conscious and explicit commonsense psychologising. That cannot suffice to settle the dominance issue. Proponents of the 'broad' view maintain that a great deal of folk psychologising goes on at an implicit, sub-conscious level. As noted, Bermúdez adopts a twofold strategy of opposition, partly questioning the idea that commonsense psychology could operate at an implicit level (pp.178-185 and pp.194-198), and partly proposing alternative processes of social cognition that do not involve commonsense psychology (pp.198-205). It should be clear that the preliminary clarification frames the discussion in a way which affects the first part of this strategy. It is entirely fair for Bermúdez to point out that the broad view, according to which exercises in commonsense psychology are merely the verbally projected upper volume of the mindreading iceberg, involves 'an empirical hypothesis that brings with it a considerable theoretical commitment, namely, to explain the nature of our *implicit knowledge* of commonsense psychology' (p.181). Yet once we appreciate the dependence of language-based propositional attitude psychology upon mindreading capacities which develop early, but which adults do not lose, this ceases to look like an objection: it really is a theoretical commitment which should be taken on.

There is a further reason why we should find nothing implausible in the idea of implicit mindreading. Bermúdez is ready to stress the enhancement of cognition provided by representation in the format of natural language. But he neglects the dependence of natural language upon mindreading. There is a complicated story of interdependence to be unravelled concerning the relation between language and mindreading. For it may well be proposed that our ability to detect false belief and deception has an evolutionary dependence upon communication, because using testimony as a source of belief exposes us to a new order of cognitive risk. (Sperber has emphasised this point: Sperber, 2001.) But the use of language for communication requires mindreading, if only because literally encoded meaning falls so far short of what an audience needs in order to grasp the speaker's meaning (see Sperber and Wilson, 2002, on this point). The pragmatic aspects of communication require implicit mindreading to be at work: as evidenced by the fact that we know this is an area in which the autistic encounter serious problems both in terms of comprehension and production. This suggests that we engage in a great deal of implicit mindreading of which we are not fully aware at a conscious level. By contrast, that commonsense psychology as such may not operate implicitly would not be so surprising, given the way it is characterised in terms of explanations invoking propositional attitudes.

What of the second part of the strategy? Have we ways of handling problems of social interaction which do not involve commonsense psychology (or mindreading)? Here we should agree with Bermúdez that we clearly do. We can respond appropriately to others without making any folk psychological predictions concerning behaviour. Also we can often predict behaviour in other ways. After all, even such devices as timetables help us cope with problems of social co-ordination. But does the frequency with which we use commonsense psychology settle the issue of 'domination'? And why should we care?

Our reliance on other predictive and reactive mechanisms can hardly show that commonsense psychology, combined with more fundamental forms of mindreading, is not our dominant form of social cognition. Some of the processes mentioned by Bermúdez (particularly the frames and routines associated with particular social situations and roles) may just be short-cuts which we find convenient and less demanding, especially in large-scale societies in which we frequently encounter strangers. By analogy, a moral consequentialist might be committed to act utilitarianism and yet might follow rules of conduct most of the time, supposing that particular applications of such rules could be justified, if one went to the trouble of doing so. In order to establish the dominance claim one would need to consider how a subject would proceed in a situation in which some non-mindreading mechanism for social interaction produced a different result from a prediction of commonsense psychology.

There is, however, an issue to which the sheer frequency of predictions delivered by commonsense psychology is pertinent. This concerns a line of argument which is a sort of scruffy relative of Putnam's Miracle Argument in support of realism about scientific theories. In philosophy of science the argument is advanced that we ought to believe in the truth (or at least the approximate truth) of the theories of modern science — and therefore in the forces and entities postulated by those theories — because otherwise the explanatory and, in particular, predictive successes of science would be inexplicable. Similarly it can be argued that commonsense psychology must be substantially correct — and so we should acknowledge what it posits: i.e., there really are such states as beliefs and desires — because otherwise we would not be able to cooperate and coordinate social interaction as well as we do.

This Argument from Successful Social Interaction has been quite influential. Bermúdez quotes a typical formulation by Braddon-Mitchell and Jackson (Braddon-Mitchell and Jackson, 1996) on p.180: 'The fact that we can make the predictions shows that we have cottoned on to the crucial regularities — otherwise our predictive capacities would be a miracle.' But this is not convincing. The sort of success which might only be explicable in terms of the approximate truth of a theory would seem to be predictive success. Does commonsense psychology really go in for *prediction* very much? It is certainly used for explaining after the act. Political pundits apart, one does not often hear people stating in advance how others will behave. Even if we allow expectations to count as predictions (despite the obvious fact that science could hardly exist if scientists had a habit of keeping predictions to themselves) the argument is weak, in part because Bermúdez is right that there are other ways of successfully handling social interaction. One might also add that any theory can be endorsed if only its successes are counted, while neglecting awkward surprises — though it is well known that there are a number of these for commonsense psychology (such as those considered in Doris, 2002).

The asymmetry between explanation and prediction should be stressed, and the reason why such an asymmetry is not surprising should be appreciated. For any system in which there are numerous and variable causal factors at work predictive success is hard to achieve, as John Stuart Mill pointed out long ago (Mill, 1843/1974: his examples were tidology and meteorology). This should lead us to expect explanation to be more common in folk psychology than prediction, since psychological causal factors appear to be numerous and volatile. That explanations are easier to supply is due to the fact that the occurrence of the explanandum draws our attention to the detection of relevant antecedents — obviously in general a less demanding task than monitoring potentially relevant antecedents for purposes of prediction.

Furthermore, in order to explain why someone acted in one way rather than another we need to cite the psychological causes that made for that difference. Throughout philosophical treatments of commonsense psychology it seems to be generally assumed that explanation (at any rate, causal explanation) requires subsumption under generalizations, and Bermúdez's discussion of the interface problem clearly inherits this assumption. But this is to found accounts of a folk theory upon a positivist model of science (Hempel's Deductive-Nomological Model of scientific explanation). I cannot here go into all the reasons why this is inappropriate. So I will restrict myself to suggesting that the contrastive account of causal explanation advanced by Peter Lipton (Lipton, 1990, 1991) provides an alternative way of construing commonsense psychological explanations that deserves to be explored.

There is unfinished business here concerning the nature of commonsense psychological explanation. What is pertinent to our issue is that it has been too readily assumed that if such explanation is causal it must involve generalizations. But this assumption seems to be based upon a model of explanation which was only ever advanced as a scientific ideal, and to which there are known alternatives in the philosophy of science. Whatever else we conclude from this situation, it does at least suggest we should be wary of claims based upon a philosophical understanding of commonsense psychology. Bermúdez relies on a standard philosophical characterisation, offering the following typical summary:

> 'Commonsense psychology is an explanatory tool that explains and makes sense of behaviour by interpreting it as the result of beliefs, desires and other propositional attitudes.' (pp.320-1)

My main objection is that this is too shallow an account, based upon verbally communicated exercises, and that it fails to do justice to the extent to which our capacities in folk psychology both developmentally depend upon, and also operationally continue to involve, more basic mindreading capacities. A pre-emptive piece of philosophical clarification only serves to entrench this neglect in Bermúdez's position, and it is really this which creates the impression that his overall architecture is exemplified in the domain of mindreading.

Let us turn, finally, to the question of overall cognitive architecture. If I have undermined the support Bermúdez draws from consideration of the interface problem, this does not show his proposal is wrong. One feature of Bermúdez's position is that he can allow cognitive processing to be more varied in kind than either the Fodorian 'Divided Mind' (peripheral modularity plus non-modular central process) or Massive Modularity hypotheses. So perhaps it does deserve to be regarded as a reasonable synthesis, rather than a Tychonic compromise?

At this level, when considering the core theories which inform research programmes, completely decisive arguments are not likely to be available and we should instead be guided by strategic considerations, in particular as to which position is likely to be the most progressive. In my opinion, Bermúdez accords too much weight to Fodor's Problem — the problem of explaining how modular systems could possibly produce the flexibility and creativity that we find in human cognition. Or rather I should say that he weights the problem in the wrong direction. For Fodor's Problem is an issue which advocates of Massive Modularity need to address; and indeed they have made a start on attempting to deal with this problem (Carruthers, 2003, 2005; Sperber, 2005). But it would be wrong to think that the undeniable difficulty of dealing with this problem counts against Massive Modularity. In other words, it is very much an anomaly which needs to be tackled, rather than a refutation. And surely there must be some way of tackling the anomaly presented by the fact that human reason appears to be, as Descartes once put it, a 'universal instrument'. It seems to me healthier to address this problem within the massively modular framework than to maintain that in some respects human cognition is domain-general and hence non-modular. That does not contribute to explaining how human thought can be flexible and creative, nor does it test whether there may be some limit to the universality of human reason.

That is one of the reasons why the massively modular research programme seems to hold out more promise of being progressive than either Fodor's model or Bermúdez's synthesis. A difficulty that Bermúdez incurs is that, in his framework, the concepts of commonsense psychology only apply in a full sense to those who have a propositional attitude complex; and since the propositional attitude complex is taken as operating in a natural language format this makes the attribution of beliefs and knowledge to infants and animals problematic. This seems to me a serious difficulty. But as his summary indicates a willingness to tackle this as an anomaly it might be less than fair to treat this as a weakness in his position.

Instead I would prefer to highlight certain 'interface problems' which Bermúdez shares with the Massive Modularity Hypothesis. One of these is the interface between representations in the format of natural language and the inputs and outputs of domain-specific modules. A consensus is emerging that natural language can serve as an inter-modular *lingua franca*, enabling information processed by different modules to be combined. That still leaves us with a problem of explaining how further cognitive processing is then possible: what feeds off the stream of linguistic representations? Another interface problem arises as soon as one allows Darwinian modules in addition to Fodorian modules. Fodorian input modules are fixed to fire automatically in response to the output of sensory transducers. But how is input channelled into Darwinian modules, given that it could come from a wide range of sources? Fodor (Fodor, 2000) tries to use this as a proof that there cannot be any such modules, and that therefore the Massive Modularity Hypothesis must be wrong. Since Bermúdez accepts that there are Darwinian modules he has to resist this argument (see pp.239-240). If the point falls short of a refutation, it has to be allowed that this is a tough problem which demands further attention.

But what of *the* interface problem? The development of natural science generated worries over how the world according to science relates to the world as manifest to our experience. The development of cognitive science seems set to recapitulate this issue of integration in an even more intimate way, leading us to question whether its theories can accord with our own self-image. However, we have not really resolved the issue about how the world according to physics interfaces with the world as we experience it, despite repeated attempts from John Locke (on primary and secondary qualities) onwards. So the philosophical attitude towards the interface problem should be more patient, and much more modest concerning our knowledge of commonsense psychology. I have argued that what we take to be commonsense psychology is largely a practice founded upon underlying mentalistic cognition. We need to work out which parts are direct projections from our underlying mental cognition and which parts are later societal additions. So in particular we need a better understanding of cultural variation and pancultural uniformity in commonsense psychology. There is a growing literature in this area (e.g., Vinden, 1996; Lillard, 1999; Knight et al., 2004; Callaghan et al., 2005). One strand in the current AHRC project on *Culture and the Mind*, aims to contribute to this with the aid of anthropological researchers based in cultures with wide geographical dispersion. This is no small field of inquiry. At this stage it is probably best to admit that we do not know very much about commonsense psychology.

References

Baron-Cohen, S., Cox, A., Baird, G., Swettenham, J., Drew, A., Nightingale, N., Morgan, K. and Charman, T. 1996. "Psychological Markers of Autism at 18 Months of Age in a Large Population." *British Journal of Psychiatry*, 168: 158-163.

Baron-Cohen, S. and Swettenham, J. 1996. "The Relationship between SAM and ToMM: Two Hypotheses." In Carruthers, P. and Smith, P.K. eds. *Theories of Theories of Mind*. Cambridge University Press.

Baron-Cohen, S., Wheelwright, S., Cox, A., Baird, G., Charman, T., Swettenham, J., Drew, A. and Doehring, P. 2000. "Early Identification of Autism by the Checklist for Autism in Toddlers

(CHAT).” *Journal of the Royal Society of Medicine* 93: 521-525.

Bermúdez, J.L. 2005. *Philosophy of Psychology: A Contemporary Introduction*. London: Routledge.

Braddon-Mitchell, D. and Jackson, F. 1996. *Philosophy of Mind and Cognition*. Oxford: Blackwell.

Callaghan, T., Rochat, P., Lillard, A., Claux, M.L., Odden, H., Itakura, S., Tapanya, S., and Singh, S. 2005. “Synchrony in the Onset of Mental-State Reasoning: Evidence from Five Cultures.” *Psychological Science* 16: 378-384.

Carruthers, P. 2003. “On Fodor’s Problem.” *Mind & Language* 18: 502-523

Carruthers, P. 2005. “Distinctively Human Thinking: Modular Precursors and Components.” In P. Carruthers, S. Laurence and S. Stich eds. *The Innate Mind: Structure and Contents*, Oxford University Press: 69-88.

Clements, W.A. and Perner, J. 1994. “Implicit Understanding of Belief.” *Cognitive Development*, 9: 377-395.

Clements, W.A. and Perner, J. 2001. “When Actions Really Do Speak Louder Than Words? But Only Implicitly: Young Children’s Understanding of False Belief in Action.” *British Journal of Developmental Psychology* 19: 413-432.

Doris, J. 2002. *Lack of Character: Personality and Moral Behavior*. Cambridge University Press.

Fodor, J.A. 1983. *The Modularity of Mind.* Cambridge, MA: MIT Press.

Fodor, J.A. 2000. *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.

Knight, N., Sousa, P., Barrett, J. and Atran, S. 2004. “Children’s Attributions of Beliefs to Humans and God: Cross-Cultural Evidence.” *Cognitive Science* 28: 117-126.

Lillard, A. 1999. “Developing a Cultural Theory of Mind: The CIAO Approach.” *Current Directions in Psychological Research* 8: 57-61.

Lipton, P. 1990. “Contrastive Explanation.” In D. Knowles ed. *Explanation and Its Limits*. Cambridge: Cambridge University Press.

Lipton, P. 1991. *Inference to the Best Explanation*. London: Routledge.

Mill, J.S. (1843/1974) *A System of Logic Ratiocinative and Inductive Books IV-VI*, in Robson, J.M. ed. *Collected Works of John Stuart Mill Volume VIII*. London: Routledge.

Onishi, K.H. and Baillargeon, R. 2005. “Do 15-Month-Old Infants Understand False Beliefs?” *Science* 308: 255-258.

Sperber, D. 2001. “An Evolutionary Perspective on Testimony and Argumentation.” *Philosophical Topics* 29: 401-413.

Sperber, D. 2005. “Modularity and Relevance: How Can a Massively Modular Mind Be Flexible and Context-Sensitive?” In P. Carruthers, S. Laurence and S. Stich eds. *The Innate Mind: Structure and Contents*, Oxford University Press: 53-68.

Sperber, D. and Wilson. D. 2002. “Pragmatics, Modularity and Mind-reading.” *Mind & Language* 17: 3-23.

Vinden, P. 1996. "Junin Quechua Children's Understanding of Mind." *Child Development* 67: 1707-1716.

In Defence of the Autonomous Mind

E. J. Lowe

Department of Philosophy
University of Durham

José Bermúdez's *Philosophy of Psychology: A Contemporary Introduction* (Bermúdez 2005) provides a wonderfully clear and well-informed survey, analysis and critique of current views in an important area of intersection between philosophical and scientific thought. Its organizing theme is that of the four influential 'pictures of the mind' that, according to Bermúdez, inform contemporary research into and reflection on the psychological domain. I have no doubt that he is correct in identifying these pictures — the autonomous mind, the functional mind, the representational mind, and the neurocomputational mind — as being the four leading paradigms in this field and I can agree with much that he says in criticism of various aspects of them. He is surely right to attempt, as he does, to look towards a fifth picture that aims to synthesize the most promising features of the previous four. A fuller working-out of that fifth picture is something that all workers in the field will look forward to with keen interest. However, Bermúdez would not thank me for confining my remarks on his excellent book to ones expressive of bland praise. First-rate philosopher as he is, he will be looking for challenges to his views and defences of the views that he himself challenges. I shall try to oblige by saying something in defence of the picture of the mind that Bermúdez seems to find least satisfactory of all, that of *the autonomous mind*.

1 The Autonomous Picture of the Mind

It is my belief that there is much more to be said in favour of the autonomous picture of the mind than Bermúdez seems to allow. Straightaway, however, I should lay my cards on the table and confess that my sympathies in the philosophy of mind lie largely with those of a dualist persuasion (see, especially, Lowe 1996). This is not at all to say that I regard myself as a neo-Cartesian dualist, but I do consider that there are strong reasons for denying that mental states are identical with or even 'realized by' neurophysiological states and for contending that causal explanation by reference to mental states is not reducible to or eliminable in favour of causal explanation by reference to neurophysiological states. I do not, however, want to try simply to sidestep or evade what Bermúdez calls 'the interface problem': 'How does commonsense psychological explanation interface with the explanations of cognition and mental operations given by scientific psychology, cognitive science, cognitive neuroscience and the other levels in the explanatory hierarchy?' (Bermúdez 2005: 35). Certainly, I believe that psychophysical dualists like myself need to offer some coherent account of how causal explanation in terms of mental states meshes with causal explanation in terms of neurophysiological states, rather than airily proposing that the two levels of explanation pass each other by like ships in the night.

So what, according to Bermúdez, is the picture of the mind that he dubs that of 'the autonomous mind'? He asserts that 'The key tenet of the autonomous conception of the mind is that there is a radical incommensurability between the type of explanation at play in commonsense psychology and that involved in explanation at the subpersonal level' (Bermúdez 2005: 52). However, this should already indicate, in the light of my previous remarks, that Bermúdez and I do not see entirely eye to eye on how to characterize the autonomous conception of the mind. I concede that there are proponents of this conception who would readily agree with Bermúdez's characterization of its 'key tenet', including the three philosophers whom he mentions most often in this connection — Donald Davidson, Jennifer Hornsby, and John McDowell. However, if the kind of naturalistic psychophysical dualism that I myself favour is to be assigned to any of Bermúdez's four pictures of the mind, it must surely be assigned to the autonomous conception of the mind — and yet, I do not altogether agree with Bermúdez's statement just quoted above. This is despite the fact that my self-confessedly dualist position is, if anything, more extreme even than what Bermúdez calls the 'more extreme version of the autonomy theory, associated with John McDowell and Jennifer Hornsby, (which] denies the claim of token-identity [between mental and neurophysiological events] characteristic of [Davidson's] anomalous monism' (Bermúdez 2005: 52), since neither of those philosophers, I believe, would accept the epithet 'dualist' as applying to themselves. More precisely, while I do believe that causal explanation by reference to the mental states of human persons is importantly *different* from causal explanation by reference to their subpersonal neurophysiological states, I consider it a vital part of any adequate theory of mind to attempt to show that and how such explanations not only are not in competition with each other, but also are mutually supportive and complementary.

Here I should perhaps emphasize that, although my defence of the autonomous mind will draw extensively on considerations concerning the truth and falsity of counterfactual conditionals relating to mental and neurophysiological states, it is by no means the case that I want to appeal to what Bermúdez himself calls 'the counterfactual approach' to causation, which he opposes to the nomological or law-based approach and which he finds problematic for various reasons (see Bermúdez 2005: 163–70; that I share some of his doubts may be gathered from Lowe 2002, ch. 10). In fact, I shall appeal to no particular theory of causation at all, but simply rely on the fact that any plausible theory of causation must at least concede that there is an intimate logico-semantic relationship between causal statements and counterfactual conditionals, of such a kind that, very often, we may unproblematically identify a counterfactual conditional whose truth or falsity is implied by the truth or falsity of a given causal claim.

2 Two Different Perspectives on the Causal Explanation of Voluntary Action

In order to keep matters relatively simple and to confine my discussion to manageable proportions, I shall concentrate on issues concerning voluntary and deliberative human action, where it is most obviously pressing that some coherent story needs to be told as to how mental and neurophysiological causes interrelate with one another — that is, where the 'interface problem' is particularly acute. So let us focus on a specific case of such an action, such as an agent's deliberate (that is, premeditated and entirely voluntary) raising of an arm, for whatever reason (for instance, in order to catch a lecturer's attention with a view to asking a question). Now, what seems relatively uncontroversial, on the purely neurophysiological side of the causal story involved in such a case, is that if we were to trace the purely *bodily* causes of the relevant peripheral bodily event — in this case, the upward movement of the agent's arm on the given occasion — backwards in time indefinitely far, we would find that those causes *ramify*, like the branches of a tree, into a complex maze of antecedent events in the agent's nervous system and brain — many of the neural events in the agent's brain being widely distributed across fairly large areas of the motor cortex and having no single focus anywhere, with the causal chains to which they belong possessing, moreover, no distinct *beginnings* (see, e.g., Deecke, Scheid & Kornhuber 1969 and Popper & Eccles 1977: 282 ff. and 293 f.). And yet, intuitively, the agent's mental act of *decision* or *choice* to move the arm would seem, from an introspective point of view, to be a *singular* and *unitary* occurrence which somehow *initiated* his or her action of raising the arm. The immediate question, then, is how, if at all, can we reconcile these two apparent facts? It seems impossible to *identify* the agent's act of choice with any individual neural event, nor even with any combination of individual neural events, because it and they seem to have such different causal features or profiles. The act of choice seems to be unitary and to have, all by itself, an 'initiating' role, whereas the neural events seem to be thoroughly *disunified* and merely to contribute in different ways to a host of different ongoing causal chains, many of which lead independently of one another to the eventual arm-movement.

I believe that a psychophysical dualist version of the autonomous conception of the mind can enable us to see how *both* of these causal perspectives on deliberative physical action can be correct, without one being reducible to the other and without there existing any sort of rivalry between the two. First of all, the act of choice is attributable to the *person* whereas the neural events are attributable to parts of the person's *body*: and a person and his or her body are, according to this conception of the mind, *distinct* things, even if they are not *separable* things (compare Baker 2000). Moreover, the act of choice *causally explains* the bodily movement — the upward movement of the arm — in a different way from the way in which the neural events explain it. The neural events explain why the arm moved *in the particular way* that it did — at such-and-such a speed and in such-and-such a direction at a certain precise time. By contrast, the act of choice explains why a movement *of that general kind* — in this case, a rising of the agent's arm — occurred around about the time that it did. It did so because shortly beforehand the agent decided to raise that arm. The decision certainly did not determine the precise speed, direction, and timing of the arm's movement, only *that* a movement of that general sort would occur around about then. The difference between the two kinds of causal explanation reveals itself clearly, I suggest, when one contemplates their respective *counterfactual* implications. If the agent had not decided to raise his or her arm, there wouldn't have been an arm-movement of that kind *at all* — the arm would either have remained at rest or, if the agent had decided to make another movement instead, it would have moved in a quite different way. It doesn't seem, however, that one can isolate any neural event, or any set of neural events, whose non-occurrence would have had *exactly the same consequences* as the non-occurrence of the agent's decision. Rather, the most that one can say is that if this or that neural event, or set of neural events, had not occurred, the arm-movement might have proceeded in a somewhat different manner — more jerkily, perhaps, or more quickly — *not* that the arm would have remained at rest, or would instead have moved in a quite different kind of way.

3 A Counterfactual-based Argument Against Psychoneural Causal Identity

This last point is an extremely important one and requires further elucidation. As Bermúdez himself acknowledges, it is now standard practice amongst philosophers of logic and language to interpret counterfactual conditionals in terms of possible worlds, very roughly as follows (see, especially, Lewis 1973, although I do not replicate every detail of his account, but only those that are germane to the issues now under discussion). A counterfactual of the form 'If it were the case that *p*, then it would be the case that *q*' is said to be true if and only if, in the *closest* possible world in which *p* is the case, *q* is also the case — where the 'closest' possible world in question is the one in which *p* is the case but otherwise *differs minimally* from the actual world. Now, suppose that a physicalist in the philosophy of mind were to propose that the agent's decision, *D*, to raise his or her arm on a given occasion — the agent's mental act of choice — is identical with a certain neural event, *N*, which is correctly identifiable as being a *cause* of the subsequent bodily event, *B*, of the arm's rising. (Here I must stress that *D*, *N*, and *B* are, each of them, supposed to be *particular events*, each occurring at a particular moment of time, with *B* occurring at least an appreciable fraction of a second later than *D* and *N*, since our decisions to act do not take effect immediately — and the physicalist must suppose, of course, that *D* and *N* occur at the *same* time, since he holds them to be identical. And let me add, too, that I do not wish to get embroiled here in the debates concerning Benjamin Libet's celebrated but highly controversial experiments on the precise timing of volitions (Libet 1985), as this would sidetrack me from my present concerns.) Let us concede, consequently, that the following counterfactual is true: 'If *N* had not occurred, then *B* would not have occurred'. All that I am presupposing here is that if *N* was indeed a cause of *B*, then the foregoing counterfactual is true. The physicalist cannot, I think, have any quarrel with me on this account. I am not taking any advantage, then, of the various reasons that have been advanced for doubting, at least in some cases, whether causal statements entail the corresponding counterfactuals (for discussion of which see Lowe 2002, ch. 10). What I am now interested in focusing on is the following question: what sort of event *would* have occurred, instead of *B*, if *N* had not occurred? In other words: in the closest possible world in which *N* does not occur, what sort of event occurs instead of *B*? My contention is that what occurs in this world is an event *of the same sort as B*, differing from *B* only very slightly. The reason for this is as follows.

It seems evident, from what we know about the neural causes of an event such as *B*, that *N* must be an *immensely complex* neural event: it must be, in fact, the sum (or 'fusion') of a very large number of individual neural events, each of them consisting in some particular neuron's firing in a particular way. (Recall, here, that *N* must be supposed to occur an appreciable amount of time *before B*, at a time at which the neural antecedents of *B* are many and quite widely distributed across the agent's cerebral cortex.) It would be utterly implausible for the physicalist to maintain, for example, that the agent's decision *D* is identical with the firing of just a *single* neuron, or even of a small number of neurons. If *D* is identical with any neural event *at all*, it can surely only be identical with an extremely complex one, consisting in the firing of many neurons distributed over quite a large region of the agent's cerebral cortex. However, it seems indisputable that if *N* is, thus, the sum of a very large number of individual neural events, then the *closest* world in which *N* itself does not occur is a world in which *another* highly complex neural event, *N**, occurs, differing *only very slightly* from *N* in respect of the individual neural events of which it is the sum. In other words, *N** will consist of *almost exactly the same* individual neural events as *N*, plus or minus one or two. Any possible world in which a neural event occurs that differs from *N* in *more* than this minimal way simply will not qualify as the *closest* possible world in which *N* does not occur. This is evidently what the standard semantics for counterfactuals requires us to say in this case. But, given what we know about the functioning of the brain and nervous system, it seems clear that, in the possible world in which *N** occurs, it causes a bodily event *very similar* to *B*, because such a small difference between *N* and *N** in respect of the individual neural events of which they are respectively the sums cannot be expected to make a very big difference between their bodily effects. There is, we know, a good deal of redundancy in the functioning of neural systems, so that the failure to fire of one or two motor neurons, or the abnormal firing of one or two others, will typically make at most only a minimal difference with regard to the peripheral bodily behaviour that ensues. Thus, the answer to the question posed earlier — what sort of bodily event would have occurred instead of *B*, if *N* had not occurred? — is this: a bodily event *very similar to B*. In other words, if *N* had not occurred, *the agent's arm would still have risen in almost exactly the same way as it actually did*.

Now, I hope, we can see the importance of this conclusion. For, if we ask what sort of bodily event would have occurred instead of *B* if *the agent's decision*, *D*, to raise his or her arm had not occurred, then we plausibly get a very different answer. Very plausibly, if *D* had not occurred — if the agent had not made the very act of choice that he or she did to raise the arm — then the arm *would not have risen at all*. It is, I suggest, quite incredible to suppose that if the agent had not made *that* very decision, *D*, then he or she would have made another decision virtually indistinguishable from *D* — in other words, *another* decision to raise the arm in the same, or virtually the same, way. On the contrary, if the agent had not made *that* decision, then he or she would either have made a quite different decision or else no decision at all. Either way — assuming that there is nothing defective in the agent's nervous system — the arm *would not* have risen almost exactly as it did.

If all of this reasoning is correct, then it follows unavoidably that the decision *D* cannot be identical with the neural event *N* with which the physicalist proposes to identify it, for the counterfactual implications of the non-occurrence of these two events are quite different. If *D* had not occurred, the agent's arm would not have risen at all, but if *N* had not occurred, it would have risen almost exactly as it did. The ultimate reason for this — according to the autonomous conception of the mind that I favour — is that a mental act of choice or decision is, in a strong sense, a *singular* and *unitary* event, unlike a highly complex sum or fusion of independent neural events, such as *N*. *N** differs from *N* only in excluding one or two of the individual neural events composing *N* or including one or two others. That is why *N* and *N** can be so similar and thus have such similar effects. But *D* cannot intelligibly be thought of, in like manner, as being *composed* of myriads of little events and that is why, in the closest possible world in which *D* itself does not occur, there does *not* occur another decision *D** which differs from *D* as little as *N** differs from *N*. (I should add that, although I do not have space enough to demonstrate this in detail here, the foregoing line of argument sustains not only the conclusion that the mental and neural causes of voluntary bodily movements must be numerically *distinct*, but also the stronger conclusion that those mental causes cannot even be taken to be 'realized by' any of those neural causes — where 'realization' is taken to be a relation distinct from identity itself, in virtue of which realized events or states inherit their causal features entirely from those of the events or states that realize them.)

4 Intentional Causation Versus Physical Causation

So far, I have tried to explain why the mental and neural causes of voluntary bodily movements must be distinct, consistently with allowing, as I do, that such movements have *both* mental *and* neural causes. Now I want to say a little more about the respects in which mental causation is distinctively different from bodily or physical causation. Most importantly, then, mental causation is *intentional* causation — it is the causation of an *intended* effect *of a certain kind*. Bodily causation is not like this. All physical causation is 'blind', in the sense that physical causes are not 'directed towards' their effects in the way that mental causes are. *Both* sorts of causation need to be invoked, I believe, in order to give a full explanation of human action and the autonomous conception of the mind seems best equipped to accommodate this fact. The very *logic* of intentional causation differs, I venture to say, from the *logic* of bodily causation. Intentional causation is *fact* causation, while bodily causation is *event* causation (for more on this distinction, see Bennett 1988 and also Lowe 2002, ch. 9). That is to say, a choice or decision to move one's body in a certain way is causally responsible for the *fact* that a bodily movement *of a certain kind* occurs, whereas a neural event, or set of neural events, is causally responsible for a *particular* bodily movement, which is a particular *event*. The decision, unlike the neural event, doesn't causally explain why that *particular* bodily movement occurs, not least because one cannot *intend* to cause a particular future event, only to bring it about that an event of a certain kind occurs. (One can only intend something if one can make it *an object of thought*: but, very plausibly, one cannot make an *as yet non-existent* future event the object of one's thought — one can at most think of the future as including *an* event of a certain kind, such as a rising of one's arm, at a certain time or within a certain period of time.)

As I have just implied, the two species of causal explanation, mental and physical, are both required and are mutually complementary, for the following reason. Merely to know why a *particular* event of a certain kind occurred is not necessarily yet to know why an event of *that* kind occurred, as opposed to an event of some other kind. Intentional causation can provide the latter type of explanation in cases in which bodily causation cannot. More specifically: an event, such as a particular bodily movement, which may appear to be merely *coincidental* from a purely *physiological* point of view — inasmuch as it is the upshot of a host of independent neural events preceding it — will by no means appear to be merely coincidental from an *intentional* point of view, since it was an event *of a kind* that the agent intended to produce (see further Lowe 1999).

Notice, here, that the aforementioned fact — that a mental decision, *D*, to perform a certain kind of bodily movement, cannot be said to cause the *particular* bodily event, *B*, of that kind whose occurrence renders that decision successful — is already implied by the argument that I developed a little while ago. For, given that *D* is *not identical* with the actual neural cause, *N*, of *B*, the closest possible world in which *N* does not occur *is still a world in which D occurs* — but in that world a slightly different bodily movement, *B**, ensues, being caused there by a slightly different neural cause, *N**. (Clearly, if *D* is not identical with *N*, then there is no reason to suppose that the closest world in which *N* does not occur is also one in which *D* does not occur, for a world in which *both* of these events do not occur evidently differs more from the actual world than a world in which just *one* of them does not occur, other things being equal.) However, this means that the occurrence of *D* is compatible with the occurrence of two *numerically different* bodily movements of the same kind, *B* and *B**, and hence does not causally determine *which* of these occurs, but only that *some* bodily movement of their kind occurs.

At this point, I anticipate the following possible objection on the part of the physicalist. Couldn't the physicalist simply *concede* that the complex neural event *N*, in our example, is not identical with the mental decision *D* — and thereby concede that *D* does not cause the *particular* bodily movement, *B*, that is caused by *N* — while still insisting that *D* is identical with *some* neural event, call it *M*, which has precisely the causal role that I am attributing to *D*? According to this view, *D* is identical with a neural event, *M*, which causally explains why *some* bodily movement of *B*'s kind occurred, but not why *B* in particular occurred. No — such a position is not tenable, for reasons which we have already encountered. Recall that I argued that the following counterfactual conditional is true: 'If *D* had not occurred, then *no* bodily movement of *B*'s kind would have occurred'. That is to say, if the agent had not performed that decision to raise his or her arm, then the arm would not have risen in anything like the way that it did — it would either have moved in some quite different way, or not at all, because if the agent had not made *that* decision, he or she would either have decided to do something quite different or else not have decided to do anything. Can the same thing be said with regard to the putative neural event *M*? No, it can't. This is because, once again, plausibility demands that the physicalist takes *M* to be an *extremely complex* neural event, composed of the firings of very many individual neurons, so that the closest possible world in which *M* itself does not occur will be one in which a neural event, *M**, occurs which differs from *M* only in respect of the firing of one or two individual neurons. And it simply isn't credible to suppose that this very small difference between *M* and *M** should make all the difference between the agent's arm rising and some quite different kind of bodily movement occurring. Consequently, the counterfactual conditional that is true of *M* is this: 'If *M* had not occurred, then a bodily movement of *B*'s kind would still have occurred'. So, once more, because different counterfactuals are true of *D* and *M*, *D* and *M* cannot be identical. The physicalist's new proposal encounters exactly the same difficulty as did his original proposal. The difficulty is that mental causes like *D* have a *strong unity* which fails to characterize extremely complex neural events such as *N* and *M*. Because of this lack of strong unity, the closest worlds in which events like *N* and *M* do not occur are worlds in which the vast majority of *their parts* still occur, with the consequence that similar bodily effects still ensue.

5 Reasons, Causes, and Freedom of Action

Much more can and should be said on these matters, but since I have discussed many of them extensively elsewhere (see again, in particular, Lowe 1999), I shall rest content with the foregoing remarks for present purposes. Here, however, it may be asked: *But what about the causes of an agent's acts of decision or choice?* Are *these* bodily, or mental, or both? My own opinion is that an act of decision or choice is *free*, in the 'libertarian' sense — that is to say, it is *uncaused* (see further Lowe 2003a). This is not to say that decisions are simply *inexplicable*, only that they demand explanations of a non-causal sort. Decisions are explicable in terms of *reasons*, not causes. That is to say, if we want to know why an agent *decided* to act as he did, we need to inquire into *the reasons in the light of which* he chose so to act (compare Dancy 2000). Since decisions are, according to my version of the autonomous conception of the mind, attributable to the person and not to the person's body or any part of it, there is no implication here that any *bodily* event is uncaused. I think that Bermúdez is unduly dismissive of this sort of position, when he remarks that 'It was for a time fashionable in the 1950s and 1960s (particularly among philosophers inspired by

Wittgenstein) to argue that psychological explanations looked for the reasons for which agents performed actions, rather than the causes of those actions, but few philosophers would nowadays deny that reason-giving explanation is a species of causal explanation' (Bermúdez 2005: 53). I think that the tide has begun to turn again on this issue. It's not that I want to exclude altogether the idea of causal explanation in terms of mental states in favour of purely rational explanation in the psychological sphere — as my earlier arguments make manifest. However, I do want to help to reinstate the idea that reason-giving explanation is *not* a species of causal explanation and that it is one form of explanation that is distinctive of the psychological sphere.

But now it may be wondered: how is it really possible for mental acts of decision to explain anything in the physical domain, if that domain is *causally closed*, as many contemporary philosophers of mind — and just about all physicalists — assume? Let us consider, however, precisely how the putative causal closure of the physical domain is to be defined, for this is no simple matter (see Lowe 2000). According to one popular view (endorsed, for example, by Kim 1993), the thesis of physical causal closure amounts to the claim that no chain of causation can lead backwards from a purely physical effect to antecedent causes some of which are *non*-physical in character. But intentional causation according to the autonomous picture of the mind, as I have tried to characterize it earlier, does not violate the thesis of physical causal closure just stated, since it does not postulate that mental acts of decision or choice are events *mediating between bodily events* in chains of causation leading to purely physical effects: it does not postulate that there are 'gaps' in chains of physical causation that are 'filled' by mental events. As we have seen, according to this picture of the mind, a decision can explain the fact that a bodily movement *of a certain kind* occurred on a given occasion, but not the *particular* movement that occurred.

Even so, it may be protested that if physical causation is *deterministic*, then there is really no scope for intentional causation on the model that I am defending to explain anything physical, because the relevant counterfactuals will all simply be *false*. It will be *false*, for instance, to say that if the agent had not decided to raise his or her arm, then a rising of the agent's arm would not have occurred: rather, precisely the same bodily movement *would* still have occurred, caused by precisely the same physical events that actually did cause it — for if physical determinism is true, there was never any real possibility that those physical events should not have occurred, nor that they should have had different effects. Maybe so. But, in view of the developments in quantum physics during the twentieth century, we now know that physical causation is *not* in fact deterministic, so the objection is an idle one and can safely be ignored. The autonomous model of intentional causation that I am proposing may nonetheless still seem puzzling to many philosophers, but if so then I suggest that this will be because they are still in the grip of an unduly simple conception of what causation involves — one which admits only of the causation of one event by one or more antecedent events belonging to one or more chains of causation which stretch back indefinitely far in time. Since this seems to be the only sort of causation that is recognized by the physical sciences, intentional causation on the autonomous model is bound to be *invisible* from the perspective of such a science (compare Lowe 2003b). To a physicalist, this invisibility will seem like a reason to dismiss the autonomous conception of intentional causation as spurious, because 'non-scientific'. I hope that to more open-minded philosophers it will seem more like a reason to perceive no genuine conflict between explanation in the physical and biological sciences and another, more humanistic way of explaining our intentional actions, by reference to our choices or decisions and the reasons for which we make them. Perhaps this kind of explanation might even find a place, however modest, within Bermúdez's projected fifth picture of the mind.

References

Baker, L. R. 2000. *Persons and Bodies: A Constitution View*. Cambridge: Cambridge University Press.

Bennett, J. 1988. *Events and their Names*. Oxford: Clarendon Press.

Bermúdez, J. L. 2005. *Philosophy of Psychology: A Contemporary Introduction*. London & New York: Routledge.

Dancy, J. 2000. *Practical Rationality*. Oxford: Clarendon Press.

Deecke, L., Scheid, P & Kornhuber, H. H. 1969. "Distribution of Readiness Potential, Pre-Motion Positivity and Motor Potential of the Human Cerebral Cortex Preceding Voluntary Finger Movements." *Experimental Brain Research* 7: 158–168.

Kim, J. 1993. "The Non-Reductivist's Troubles with Mental Causation." In J. Heil and A. Mele (eds.), *Mental Causation*. Oxford: Clarendon Press: 189–210.

Lewis, D. K. 1973. *Counterfactuals*. Oxford: Blackwell.

Libet, B. 1985. "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action." *Behavioral and Brain Sciences* 8: 529–566.

Lowe, E. J. 1996. *Subjects of Experience*. Cambridge: Cambridge University Press.

Lowe, E. J. 1999. "Self, Agency and Mental Causation." *Journal of Consciousness Studies* 6: 225–239.

Lowe, E. J. 2000. "Causal Closure Principles and Emergentism." *Philosophy* 75: 571–585.

Lowe, E. J. 2002. *A Survey of Metaphysics*. Oxford: Oxford University Press.

Lowe, E. J. 2003a. "Personal Agency." In A. O'Hear (ed.), *Minds and Persons*. Cambridge: Cambridge University Press: 211–227.

Lowe, E. J. 2003b. "Physical Causal Closure and the Invisibility of Mental Causation." In S. Walter & H.-D. Heckmann (eds), *Physicalism and Mental Causation: The Metaphysics of Mind and Action*. Exeter: Imprint Academic: 137–154.

Popper, K. R. & Eccles, J. C. 1977. *The Self and its Brain: An Argument for Interactionism*. Berlin: Springer.

On Psychological Explanation and the "Interface Problem"

Alfredo Paternoster

DEIS
University of Sassari, Italy

Bermudez (from now on, B.) has written a very impressive book (let me say, again!). It is a highly remarkable work both for the scope of the subject-matter and the clarity of the arguments. Many issues are discussed in a deep and effective way. Besides, I happen to be strongly sympathetic with most of B.'s views. In particular, I agree on the idea that the so-called *interface problem* is the crucial problem in the philosophy of cognition (even if, as we shall see, my assessment of the problem is slightly different). As a consequence, B.'s taxonomy of the different approaches to the discipline based on the way they address this problem seems to me very effective.

In this contribution I shall raise a pair of points about which I am less convinced. They concern (1) the relation between the philosophy of psychology and the philosophy of mind, and (2) the account of the mind outlined in B.'s concluding remarks, especially with respect to the interface problem.

1

In the first chapter B. defines the nature and the field of the philosophy of psychology. In particular, he states what he takes to be the relation between the philosophy of psychology and scientific psychology, on the one hand, and between the philosophy of psychology and the philosophy of mind, on the other. While I found his assessment of the former correct and quite effective, I am less convinced of his view of the latter.

According to B., philosophy of psychology is distinct from philosophy of mind, since (*i*) philosophy of mind has much more to do with metaphysical (in opposition to explanatory) questions, and (*ii*) philosophy of mind is more autonomous (from science) and a priori.

I think that both claims are strongly idiosyncratic. As to the first claim, although it is true that the mind/body problem is the central problem in philosophy of mind, the metaphysics of mind is so intertwined with semantic, epistemological and explanatory questions, that it seems very hard to establish a difference on this ground. Take, e.g., computational functionalism in his canonical version, the computational-representational theory of mind (CRTM). This is no more a metaphysical account than it is an explanatory picture. It is hard to make sense of the CRTM if one separates the (alleged metaphysical) thesis that mental states are computational, and ultimately physical, states picked out at the functional level from the (alleged explanatory) thesis that computational states are required to explain the causal-inferential nature of thinking. After all, it is a materialist worry that motivates the postulation of computational states as realizers of content-bearer mental states. B. would reply that one can easily offer a metaphysical solution to the mind-body problem which is, however, totally unsatisfactory from the explanatory point of view, as is the case of the token-identity theory (cf. p. 35). However, even discounting the fact that this restriction of the philosophy of mind to the mind-body problem is arbitrarily exaggerated, it seems to me that the token-identity theory is unsatisfactory even from a metaphysical point of view, since it raises some (metaphysical) puzzles that are hardly independent from the explanatory inadequacy of the theory, such as the intelligibility of the relevant notion of identity (see, e.g., Horgan & Tye 1989;

Putnam 1999), or the lack of justification of the alleged identity, which appears to be a "brute fact" (Kim 1998). Therefore, B.'s claim that the metaphysical problem and the explanatory problem «can be pursued largely independently of each other» (p. 36) seems to be quite abstract and unmotivated.

As to the second claim, I cannot see why a philosopher of mind should be committed to it. B. holds, correctly in my opinion, that philosophy of psychology involves both conceptual analysis and empirical research, and provides good arguments to this effect. Indeed, exactly the same considerations can be done for the philosophy of mind. If a philosopher of mind worked only *a priori*, for the very same reasons mentioned by B. he would hardly obtain convincing results, doing a kind of philosophy that B. himself would not entirely appreciate.

Hence, to draw a clear distinction between the two disciplines is difficult, unless one thinks that philosophy of psychology is to be narrowed to the discussion of the psychological practice – but B. explains why this is not the case. I suggest that what can at most be derived from B.'s assumptions on the domain and methodology of the two disciplines is that philosophy of psychology is *a part* of the philosophy of mind, the part that has more to do with common sense psychological explanation, on the one hand, and with scientific issues (and, as a consequence, it is more directly committed with empirical matters), on the other.

Admittedly, this is not in itself a very important issue, but we shall see in the next section that the way B. regards metaphysical aspects in the philosophy of psychology brings him to underestimate a far more substantial problem. Let us focus, then, on substantial matters.

2

As B. nicely explains, there is a hierarchy of psychological explanations, the top of which is occupied by *common sense psychology*, an horizontal explanation of the behavior of an agent. 'Horizontal' means that behavior is explained in terms of causal antecedents belonging to the same level of description. Indeed, common sense (or folk) psychology is a *personal* level explanation; beliefs, desires and actions are personal level concepts. On B.'s view, there are four different pictures of the mind according to the way each of them deals with the problem of how to link folk psychology to lower level explanations. This is what B. calls *the interface problem*, and the four pictures are the following:

> (*i*) The *autonomous* mind, which claims that there is a strong discontinuity between folk psychology and the subpersonal lower levels. To put it roughly, mental predicates do not refer to natural kinds, whereas lower level concepts do, or at least aim to do.
>
> (*ii*) The *functional* mind, according to which mental states are individuated by their causal role, and the goal of subpersonal psychology is to individuate the mechanisms which realize, or implement such roles.
>
> (*iii*) The *representational* mind, which can be described by the slogan "Functional mind + mental representations". Mental representations are the vehicles of the content of mental states. Computer programs are the metaphor of the mental.
>
> (*iv*) The *neurocomputational* mind, which rejects both the folk psychology (i.e., folk psychology is regarded as a false theory or *proto*-theory) and the computer metaphor, in favour of a brain-oriented approach.

As I said in the beginning, this way of presenting the matter seems to me quite effective. Of course, like any other taxonomy, it is not fully uncontroversial. It can be argued, for instance, that there is so much that is shared by (*ii*) and (*iii*) with respect to what distinguishes them from (*i*) and (*iv*) -- this point is acknowledged by B. himself -- that three pictures are enough. Or that the inclusion of Dennett among the supporters of the autonomous mind clashes with the characteristic antinaturalist flavour of the autonomist picture (Dennett's position is particularly hard to accommodate in B.'s classification, since it shares some aspects with each proposed model of mind. From this point of view, Dennett is not so far from the "fifth view" outlined by B., see below). Nevertheless, the taxonomy is properly justified, on the basis of very clear criteria. In particular, it is worth to note that all the four pictures take folk psychology as a cornerstone. In fact, even when folk psychology is rejected, as in the neurocomputational approach, still folk psychology is in a way a point of reference, for it is regarded as the crucial polemical target. In other words, the neurocomputational approach characterizes itself *in opposition* to the view of folk psychology as a genuine explanatory model.

That's why the interface problem is so important. Two approaches address it in a *deflationary-eliminative* way, for opposite reasons: according to the autonomous mind common sense psychological explanation is true, but there is no need to link the common sense psychological explanation to subpersonal explanations; whereas, according to the neurocomputational mind, common sense psychological explanation is false or, at least, requires drastic revision, so that there is nothing to link: subpersonal explanations are good as they stand. The other two approaches take the problem seriously, and try to deal with it in a similar (making abstraction from details) way: subpersonal levels specify the mechanisms which *realize* the top level laws: «there are systematic relations … between the nodes of the psychological network and the physical structures in the brain that serve as the nodes of the isomorphic network at the subpersonal level.» (p. 36).

Now, it should not come as a surprise that none of the pictures is fully satisfactory: each picture «works best for a limited domain.» (p. 320). In fact, each picture picks out a certain kind of mental task as the paradigm of the mental, so that the model only appropriate for that kind of task is applied across the board -- to the whole mind. B. seems to think that there is nothing in the interface problem over and above this difficulty, and suggests to search for a new picture – the so-called "fifth view" -- which combines «the insights and analyses offered by the different approaches» (*ibid*.). By contrast, I believe that the problem is deeper. I shall argue in the following that there is a distinctive difficulty in the interface problem, and that B.'s fifth view is not in a better position to work it out, though B. is right in claiming that the range of the problem turns out to be narrower.

Let me clarify, to start, what B. intends when proposes to mix the ingredients, so to speak, of the different pictures. The suggestion is not so much that of using, for instance, both neural networks and classical algorithms to model different kinds of mental processes – although he would probably agree also on this kind of ecumenism. Remember that the four pictures taxonomy is based on the way the interface problem is addressed. Hence, the fifth view has to be regarded as a synthesis of all the pictures in rather the following sense: *a*) the interface problem is taken seriously (as the pictures *ii* and *iii* do); but, although *b*) folk psychology is regarded as a true theory (as in *i*), *c*) the role of folk psychology is played down to a significant extent (in a way, as in *iv*).

The idea is that the interface problem can be solved or, at least, made more tractable, by narrowing its range: beliefs, desires and the likes are not so relevant to the explanation of behavior as it appears to be.

According to B., the role of common sense psychology is overstated in the pictures *i*, *ii* and *iii*. In fact, not every kind of behavior requires a cluster (or a pair) of cognitive states as its cause. As B. claims, there are many ways of behaving in a non-instinctual and non-reflex manner that completely bypass the propositional attitudes. Hence, we can make sense of the behavior of others without involving propositional attitudes, for «much of our understanding of other people rests upon a range of relatively simple mechanisms and heuristics that allow us to identify patterns in other's people behavior and to respond appropriately to the patterns detected.» (p. 321).

Emphasis on belief/desire as the source of action is strictly related to the endorsement of what I call the *sandwich model* of the mind, whereby perception and action are systematically mediated by the *central cognition* layer (B. calls this model "bidimensional"). Central cognition is the sandwich filling: according to the sandwich model, perception yields beliefs, and these, in turn, trigger actions. The separation between the three layers, in particular between perception and cognition, is taken to be very neat, especially if the model is coupled (as often is) with the Fodorean view of modularity, according to which only perceptual systems are modular.

Therefore, if one rejects the sandwich model, beliefs and desires may no longer be considered as the unique causes of action. There are other mechanisms, not located at the level of folk psychology, which can explain action. Among these mechanisms, B. mentions template matching, pattern recognition, the so-called Darwinian modules (such as the famous cheater detection module postulated by Cosmides & Tooby), scripts and other "dirty and quick" (or "fast and frugal") heuristics. For instance, in order to perform the action of comforting a friend, one needs not to postulate the belief that he is sorrowful and the desire of alleviating his pain: it is enough to suppose that the perception of our friend's emotional state, for instance, seeing his face, directly triggers our behavior. As an example of quick heuristic, B. mentions the application of the TIT-FOR-TAT strategy to solve the well known prisoner's dilemma. The TIT-FOR-TAT consists in the application of the two following rules: 1. Always cooperate in the first round; 2. In any subsequent round do what your opponent did in the previous round.

Therefore, in the place of the sandwich model, B. outlines a general picture of mind, based on three main features: (*a*) a (by and large) Gibsonian approach to perception; (*b*) massive modularity; (*c*) the view that some kind of thinking (namely, metarepresentational thinking) require public language. Indeed, if (*a*) is true, perception and action are not necessarily mediated by cognition; and if (*b*) is true, there is no holistic involvement of the whole belief system when one must take a decision, or plan an action. As to (*c*), it is crucially relevant to the claim that the scope of folk psychology must be significatively narrowed: if, in fact, folk psychology, as a capacity that involves metarepresentational skills, requires language, then it can be argued that folk psychology emerged long *after* the cognitive endowments underlying many social skills. To be sure, B. does not present the theses *a*, *b* and *c* as uncontroversial truths, nor as well-confirmed hypotheses (B. himself notes, for instance, that the above-mentioned phenomenon of empathic behavior based on template matching, or "affect attunement", hardly shows that there is no recourse to folk psychology at all); however, he offers some arguments for them; and, after all, the most recent researches in cognitive science are developed in this line.

Now, my point is that this proposal will not solve the interface problem. Although I am by and large sympathetic with B.'s "corrections" of standard computationalism, I do not believe that they shift the main points of the question in an appreciable way, basically for the following reason: Either common sense mental states really play a very marginal role in the explanation of behavior, and in this case standard computationalism (as it stands) is already able to deal with the interface problem, or, as I suspect, beliefs, desires and the likes are crucially relevant in order to understand ourselves and the others, and in this case B.'s strategy is doomed to fail from the start, since mental states are ineludible *explananda* for a psychological theory. We cannot do without them, if we want to make sense of ourselves as human beings. The downsizing of folk psychology is useless in order to account for personal level, unless one intends to get rid of the personal level across the board (as neurocomputationalists are somehow inclined to do), which, however, is far from being B.'s intention. Let me illustrate and argue for this view.

Notice, first of all, that the kind of solution put forward by B. has a deflationary-eliminative style. The interface problem, in fact, has been defined as the problem of vertically vindicating the common-sense psychological explanation. B.'s proposal, instead, removes from the picture, *in some cases*, common-sense psychological explanation. Hence, strictly speaking, the interface problem, as is assessed in B.'s own terms, disappears (in some cases) rather than being solved. That's very good as well, but, insofar as B. has not disqualified folk-psychology as such – just narrowed its scope -- the problem remains untouched when folk psychology is relevant. In this sense, it seems to me that no real progress has been made.

In other words, *if* the interface problem has to be considered as a serious problem, narrowing its scope makes it less pressing, but in no way less difficult. And B. appears to think that the problem is serious indeed; otherwise, why base on it the structure of the whole book? Why be dissatisfied with the functionalist-representationalist account of the problem? And why spend, in particular, so many pages against the autonomist arguments aimed to *dissolve*, rather than deal with, the problem? Hence, the crucial question is: *what* makes the interface problem so hard? Despite the pivotal role accorded by B. to the interface problem, he is not very clear on this point. Paradoxically, the problem becomes mostly manifest in the autonomist framework, the view in which the problem is yet strongly underplayed (maybe the paradox can be explained away by saying that autonomy theorists are forced to downplay the problem just because in their perspective the problem becomes untractable!). Indeed, two theses are constitutive of the autonomist picture:

(i) Mental (= personal) properties supervene neither on computational properties nor on other natural properties.

(ii) The failure of the supervenience of the mental does not raise a problem at all.

Now, I agree with B. that (*ii*) is false. Thus, if (*i*) is true, then there is a serious problem indeed. And if (*i*) is true, *no* kind of computationalism will do.

As far as I can tell, it is the *personal* nature of the psychological predicates that makes the problem hard. Even if we do not buy (*i*), still we do not clearly understand what the relation is between a personal state such as believing that *P* (or, for that matter, perceiving an *O*) and the (alleged) underlying computational state. On the one hand, there are well-known problems with the thesis that personal events are identical to computational or any other subpersonal events, even in the token-identity case (see, e.g., Putnam 1999). On the other hand, we are dissatisfied with the standard claim that the belief state supervenes on the computational state, since the thesis is too vague. Thus, the interface problem is the problem of linking personal level predicates, hopefully *all* personal level predicates, to subpersonal descriptions. The idea of substituting, for instance, a heuristic not based on belief/desire attributions (assuming, for the sake of the argument, that these heuristics do not involve psychological reasoning at all) does not affect this problem, since not only beliefs and desires, but also mental images, perceptions and emotions have some role to play in determining behavior, and these are all personal states.

B. could reply that this is *my*, *not his* assessment of the interface problem. On his view, the problem is basically to explain behavior. The dirty and fast heuristics are an explanation. The interface problem is solved because, instead of linking a subpersonal explanation to a personal explanation, behavior (which is a personal level notion) has directly been explained in subpersonal terms. For instance, we can say something like the following: *X* has performed the action *A*, say, asking for a steak, because he has performed a template matching *T* (i.e., he has performed something allowing him to recognize the man in front of him as the waiter). *T*, in turn, is realized by the algorithm so and so.

The problem with this picture is that running a template matching is not a predicate that can be attributed to an *agent*. The correct description of the situation is rather the following: *X* performed the action *A* because he was in a recognitional state *R*. *R* is realized by a template matching *T*. *We* recognize, whereas our *mind/brains* perform template-matching. Even if the recognitional state *R* is not a *belief* state – what I am absolutely prepared to concede to B. – still, there are two different levels in question: one personal and the other subpersonal, and the problem of clarifying the relation between the two levels is still there. And if the reply were that template matching is exactly the process that allows to recognize the butcher, then I wonder in which sense folk psychology is supposed to raise an interface problem: When, say, Fodor claims that believing that *P* is instantiating a LOT formula having *P* as content, why not regard this proposal as a genuine solution to the interface problem? To explicate the relation between a recognitional state and the process *T* does not seem easier than explicating the relation between a doxastic state and, say, instantiating a LOT formula (though there could be some difference -- I will go back to this point a few lines below). Either there is no interface problem in the case of belief too, or else the interface problem persists in the case of the recognitional state. If the interface problem were not conceived of as I have suggested, i.e., as the problem of linking subpersonal states with personal states, one would not understand why the problem is so difficult. After all, the functional and the representational pictures give an answer to the problem. Why does B. not consider this answer completely satisfactory?

Arguably, the apparent gap between beliefs/desires and the underlying computational states seems to be larger than the gap between perceptual/recognitional states and the underlying computational states. In fact, since it can be argued that the ascription of beliefs, desires and the likes is subjected to normative constraints, it is hard to suppose that these states supervene on computational states, whereas experiential states are arguably norms-independent. Here one can appreciate the importance of B.'s thesis that experiential states and subpersonal perceptual states both have nonconceptual content.

I tend to agree with B. on the nonconceptual nature of experiential states (with some *caveat* that it is not worth to discuss here), whereas I am quite skeptical as regard to the attribution of content to subpersonal states (see, e.g., Egan 1992, who seems to me to have successfully argued against the *individuation* of subpersonal states in terms of content). I concede that, notwithstanding the quest of subpersonal ascription of content, experiential states are more easily regarded as supervenient on computational states, but I insist that the nature of the relation between personal states and computational states remains obscure to a certain extent. How should we exactly conceive of the relation between, say, the vision of a cat, and the state of instantiating the 2½-D sketch of that cat? If we are not able to provide a clear answer to this question -- and, as far as I can tell, we are still waiting for a clear answer --, the explanatory import of the computational approach (broadly considered, not narrowed to the picture *iii*) will be poor. In which sense are mental representations explanatory, if we are not able of stating a clear relation between computational states and first-person phenomena? What is the point of postulating representations if they do not explain phenomena? The notion of supervenience is too vague as an answer to this problem.

At this point, B. might argue that I have put into the picture a somewhat obscure metaphysical question whereas the relevant issue was purely explanatory. However, as I pointed out in the previous section, the two aspects are hard to separate. After all, the interface problem can equally be described as the problem of accounting for the alleged *causal* nature of psychological explanation (and, more generally, of all our mental states). To explain how it is possible for mental states to be causally efficacious involves investigating both the nature of the relation between personal mental properties and subpersonal properties, and the nature of the relevant notion(s) of cause. Both investigations are genuinely metaphysical. It seems to me that what we ask of the scientific study of mind are (*inter alia*) answers to questions such as: what is (= what is the nature of) a desire? What is a belief? What does it mean 'to perceive'? What is a sensation, or an emotion? The computational paradigm seems to be partly inadequate in order to give this kind of answers, for reasons that have little to do with the importance of folk psychology.

Please note that nothing in these considerations have to be interpreted as a criticism of the "mixed" model of the mind proposed by B.. Quite the contrary: like B., I am convinced that the "orthodox" (Fodorean) computationalism needs to be reformulated and integrated to some extent. In particular, an important merit of B.'s strategy is that it shows that, in order to perform a variety of tasks, we use some "quick recipes". In other words, B.'s proposal is a good antidote against the recurring attitude in philosophy of mind to assume, as paradigmatic of mental activity, too high and sophisticated skills, involving conscious inferences, scientists-like procedures etc. And this has actually some impact on the issue of the relations between levels, as, if the kind of ability that we want to explain is less sophisticated, vertical explanations turn out to be easier. Let me say again, however, that this is not the heart of the interface problem. The real question is whether a different model of representationalism is in a better position to deal with the interface problem, and I think to have showed that it is not, since the problem of the relation between the psychological explanation and the computational theories brings with itself, willy-nilly, the more general problem of the relation between personal states and subpersonal, computational states. I submit that B.'s idea that the interface problem can be addressed by simplifying the kind of cognitive tasks to explain stems from his modest sensitivity to metaphysical questions, which he takes to be out of philosophy of psychology. But that it is just to push the problem to the next-door philosopher of mind.

To summarize, B.'s strategy for dealing with the interface problem (his "fifth view") consists in downplaying folk psychology. This downplaying is based on:

a) the existence of perception/action loops without the mediation of central states;

b) the massive modularity hypothesis;

c) the linguistic nature of metarepresentational thought (in the sense that language is required for that kind of thought)

It seems to me, however, that, independently of the merits of each of these theses, the interface problem remains very hard. The reason is that the gap between the personal level and the lower levels does not depend on the primacy attributed to folk psychology. Or, to put it in a slightly different way, the interface problem is in part (arguably, in a conspicuous part) a *conceptual* problem, whereas B.'s proposals are all to be evaluated on empirical grounds. The interface problem comes from the concept of person, which seems to be recalcitrant to any form of reduction. When one turns himself to study mental subsystems, he seems to lose touch with the person. Am I thereby committed to the approach of the autonomous mind? No, as I agree with B. that this approach tends to avoid the problem, rather than to cope with it. The deflationary strategy put forward by the "autonomist" scholars is poorly justified. Hence, the best we can do is to pursue our research in B.'s style, patiently combining conceptual analysis (metaphysics included!) and empirical study.

References

Egan, F. 1992. "Individualism, Computation and Perceptual Content." *Mind,* 101: 443-459.

Horgan, T., Tye, M. 1989. "A problem with the Token-Identity Theory", in M. Tye, *The Metaphysics of Mind*. Cambridge: Cambridge University Press (chap. 1).

Kim, J. 1998. *Mind in a Physical World*. Cambridge MA: MIT Press.

Putnam, H. 1999. *The Threefold Cord*. New York: Columbia University Press.

Natural Language and the Propositional Attitude Complex

Karen Shanton

Department of Philosophy
Rutgers University

1 Introduction

In *Philosophy of Psychology: A Contemporary Introduction*, José Luis Bermúdez (2005) seems to pursue two main projects. The first of these projects is to describe and assess the current state of the philosophy of psychology. He starts, in Chapter 1, by demarcating the boundaries of the field. Next, in Chapters 2-5, he introduces a problem that he takes to be central to the philosophy of psychology, i.e. the interface problem, and describes how each of the four currently dominant pictures of the mind, i.e. the autonomous, functional, representational and neurocomputational pictures, responds to it. In these chapters, he explains that the interface problem is the question of how the commonsense level of psychological explanation is related to scientific levels of explanation and shows how the four extant pictures of the mind answer this question. He also uses these chapters to identify some of the advantages and disadvantages of each picture, e.g. he suggests that the autonomous mind picture overemphasizes the scope of commonsense psychology. Finally, in Chapters 6-10, he uses the dialectic between the four pictures to introduce specific issues in the philosophy of psychology. In these chapters, he raises issues ranging from mental causation to the structure of the cognitive architecture to the relationship between thought and language and shows how the existing pictures of the mind deal (or fail to deal) with them.

The second project is to develop and defend an alternative, fifth, picture of the mind. In the concluding chapter of the book, Bermúdez (2005) notes that a general problem seems to emerge from the discussion in the preceding chapters: though each of the existing pictures of the mind is good at accounting for certain kinds of cognition, they all have difficulties with at least some other kinds. For example, though the neurocomputational picture is very good at explaining our perceptual and recognitional abilities, it's less successful at explaining logical reasoning. Bermúdez suggests that this problem can be traced to a problem with the approach these pictures take to explaining the mind. He argues that the existing pictures each identify a single, paradigmatic type of thinking and try to model all cognition on that paradigm, e.g. the neurocomputational picture tries to model all types of cognition on low-level strategies like pattern recognition. According to Bermúdez, the problem with this type of approach is that none of the paradigms comfortably accommodates all kinds of cognition; there are, therefore, at least some kinds of cognition for which each of the existing pictures fails to provide a satisfactory explanation.

The purpose of the last chapter of the book is, then, to start to develop a picture of the mind that can avoid this problem. Bermudez (2005) begins by outlining the general structure of his alternative picture. As noted above, he traces the problem with the existing pictures to the strategy of modeling all cognition on a single paradigm; he concludes, therefore, that the alternative picture should incorporate elements of multiple paradigms. This conclusion leads him to posit what he describes as a 'three-dimensional' picture of the mind. Like the picture that he describes as 'the standard view' in Chapter 8, this picture draws a distinction between peripheral perceptual and motor processing and other (personal-level) types of processing. Unlike that picture, however, his picture also draws a distinction between different types of other processing; whereas the standard 'two-dimensional' picture subsumes all other processing under a single 'central' processing heading, his three-dimensional picture divides it into a low-level network of perception-action pathways and a high-level propositional attitude complex.

Having outlined the general structure of his alternative model, Bermúdez (2005) offers a specific account of one of its parts, i.e. the high-level propositional attitude complex. He notes that there are two kinds of propositional attitudes, i.e. first-order propositional attitudes and second-order propositional attitudes, and argues that both of these kinds of attitudes are shaped (in some fundamental way) by natural language. First-order propositional attitudes, or thoughts about the world, can be understood in terms of the rewiring hypothesis; the neural rewiring that occurs when we acquire natural language "creates potential vehicles for propositional attitudes" (2005, p. 329). Second-order propositional attitudes, or thoughts about thoughts, can be traced to the mental manipulation of natural language sentences; we are able to think about our own thoughts and the thoughts of others because we can mentally manipulate the natural language sentences that express them. By the end of the concluding chapter, then, Bermúdez has sketched a picture of the mind in which peripheral perceptual and motor processes are distinguished from other, personal-level processes, these other processes are divided into high- and low-level networks or complexes and the high-level complex is understood in terms of natural language.

There are some concerns we might raise about Bermúdez's (2005) approach to the first of these projects. For example, we might challenge the scope of his characterization of the philosophy of psychology, e.g. by questioning whether we really can use empirical testability to distinguish it from the philosophy of mind, or the scope of his treatment of certain issues, e.g. by noting that he fails to include the performance error explanation of failures on the false belief task (see, for example, Fodor, 1992; Goldman, 2006; Leslie and Polizzi, 1998) in his discussion of mindreading. However, though I think that these types of concerns are worth mentioning, I don't take them to be particularly troubling. First, such scope-based concerns can be raised about almost any introductory text. Because the fields they are intended to introduce tend to be broadly ranging and widely discussed, introductory texts typically have to impose limits on the topics they address and the depth in which they address them. This means that they can't cover every topic of possible interest or provide as specific and comprehensive a discussion as texts that are wholly dedicated to a single issue; by nature, therefore, they tend to be susceptible to scope-based objections. Second, as I will explain in greater detail in §3, I think that Bermúdez's approach to the first project has a number of positive attributes that outweigh these types of concerns.

The emphasis of my critical discussion of the book will, therefore, be on Bermúdez's (2005) approach to the second project. In the next section, §2, I'll introduce empirical findings that seem to cast doubt on a particular part of that project and consider some ways in which Bermúdez might try to respond to these findings. Following this critical discussion, in §3, I'll offer some general thoughts about the book as a whole.

2 Bermúdez's Alternative Picture of the Mind

As noted in §1, in the concluding chapter of the book, Bermúdez (2005) offers an account of the high-level propositional attitude complex that traces both first- and second-order propositional attitudes to natural language; he says that the possession of first-order propositional attitudes depends on the neural rewiring that accompanies the acquisition of natural language and the possession of second-order propositional attitudes depends on the ability to mentally manipulate natural language sentences. According to this account, then, the acquisition of natural language seems to be a prerequisite for the possession of either type of propositional attitude. If Bermúdez really is committed to this understanding of the relationship between natural language and propositional attitudes, however, he also seems to be committed to the following empirical prediction: beings that lack natural language will also lack (both first- and second-order) propositional attitudes.

Recent research in nonhuman animal and developmental psychology seems to suggest, though, that this prediction doesn't hold. The consensus among researchers in these fields appears to be that nonhuman animals and very young human infants lack natural language. For example, Clifford R. Mynatt and Michael E. Doherty note that, "no one argues that any nonhuman species uses language in its natural environment. Most animals communicate, but none use language" (1999, p. 215). Similarly, Philip G. Zimbardo and Ann L. Weber (1997) point out that most language acquisition in humans occurs between eighteen months and six years of age; though younger infants might possess a few words, e.g. 'mama' and 'dada,' they don't seem to have anything resembling a complete grasp of language. Recent studies seem to suggest, however, that members of both of these groups *do* possess propositional attitudes. More specifically, the studies suggest that pre-eighteen-month-old human infants and some species of nonhuman animals are capable of at least one type of metarepresentational thought, i.e. ascribing mental states to others.

The first type of evidence for this claim is evidence of the attribution of perceptions by nonhuman animals. In studies of rhesus macaque monkeys and scrub jays, respectively, Jonathan I. Flombaum and Laurie R. Santos (2005) and N.J. Emery and N.S. Clayton (2001) found that the members of these two species seem to be capable of attributing perceptions to others. In the Flombaum and Santos study, monkeys were presented with a naturalistic food competition task. In this task, grapes were placed in front of two human 'competitors,' one of whom was physically oriented such that he could see the grape in front of him and one of whom was physically oriented such that he couldn't see the grape. The monkeys were then given the opportunity to try to steal a grape from one of the competitors. What Flombaum and Santos found was that, even when the experimental manipulation was very subtle, e.g. one competitor had a barrier covering his eyes while the other had a barrier covering his mouth, the monkeys consistently tried to steal from the competitor who couldn't see the grape.

In the Emery and Clayton (2001) study, jays were presented with a similarly naturalistic food caching task. In this task, some of the jays were allowed to cache their food stores in private while others had to cache in view of a conspecific. The jays in both conditions were then given the opportunity to retrieve the stores in private and re-cache them in either the original location or a new location. What Emery and Clayton found was that the jays that had originally cached in view of a conspecific showed a preference for re-caching in a new location while the jays that had originally cached in private were equally likely to re-cache in either the original location or a new location.[2]

The conclusion we can draw from these two studies is that the subjects of the studies were attributing perceptions to others. Because their behaviors varied with differences in the possible contents of the others' perceptions, it seems reasonable to conclude that they were attributing those perceptions to the others. If the studies show that monkeys and jays can attribute perceptions to others, though, they show that at least some species of nonhuman animals possess second-order propositional attitudes

The second type of evidence for the claim is evidence of the attribution of beliefs by nonhuman animals and pre-eighteen-month-old human infants. In studies of chimpanzees and fifteen-month-old human infants, respectively, Brian Hare, Josep Call and Michael Tomasello (2001) and Kristine H. Onishi and Renee Baillargeon (2005) found that the members of these two groups seem to be capable of attributing beliefs to others. In these studies, the researchers tested their subjects with non-verbal versions of a theory of mind task called the false belief task. The condition of interest in this type of task is the false belief condition. In this condition, subjects both see an item being moved and see that another individual, e.g. a conspecific, doesn't see the item being moved. In nonverbal versions of the task (like the current versions), researchers then use behavioral and / or eye-tracking measures to determine how subjects expect the other individual to behave with respect to the item. More specifically, they test whether subjects expect the individual to act in accordance

[2] These findings were obtained with jays that had previous experience with pilfering from other jays' caches.

with the subject's (true) belief that the item is in the new location or his own (false) belief that it is in the original location. Traditionally, displaying the expectation that the other individual will act in accordance with his own false belief has been taken as evidence that the subject is attributing a (false) belief to him. In both of the studies cited above, subjects tended to display this expectation. This seems to suggest that the subjects in these studies were attributing beliefs to others which, in turn, suggests that they possessed second-order propositional attitudes. Like the previous two studies, then, these two studies seem to show that at least some nonlinguistic beings possess second-order propositional attitudes.

Now, Bermúdez (2005) doesn't fail to recognize the possibility of this type of evidence. In fact, he explicitly acknowledges, and attempts to respond to, similar evidence in the last chapter of the book. Whether or not his responses succeed for the specific types of evidence he has in mind, though, it's not clear to me that either those responses or any other immediately obvious responses can accommodate the current evidence.

Strategies for responding to the kinds of studies we've been discussing can be divided into at least two general types. The first type [which Bermúdez (2005) doesn't discuss] is the strategy of denying that the subjects in the cited studies lacked natural language. The proponent of the natural language-based account of the propositional attitudes might try to argue that the nonhuman animals and human infants that were tested in the above studies actually did possess natural language; even if the studies show that they possessed propositional attitudes, therefore, they don't disconfirm his prediction. Of course, if they don't disconfirm his prediction, they also don't present a problem for his account.

There is, however, an obvious problem with this type of strategy. That is, given the current state of the empirical evidence, it just doesn't seem to be the case that nonhuman animals and very young human infants have natural language; as noted above, researchers in nonhuman animal and developmental psychology tend to take the available evidence to show that these beings are nonlinguistic. Of course, it's possible that new evidence might lead us to revise this conclusion, i.e. we might discover new evidence that shows that some of the beings we had previously taken to be nonlinguistic actually do possess language. Even if we find evidence for language in some beings that we had previously taken to be nonlinguistic, however, it seems implausible that we will find such evidence for all of the beings that were tested in the above studies. For example, it seems highly unlikely that we'll discover that scrub jays have natural language. Even at best, then, this strategy would probably only be able to account for some of the studies we've been discussing.

The second type of strategy [which Bermúdez (2005) does discuss] is the strategy of denying that the subjects possessed propositional attitudes. As Bermúdez shows, this general strategy can be divided into at least two substrategies. The first substrategy is what he calls the minimalist strategy. According to this substrategy, we can simply "refus[e] to take at face value the explanatory practices of cognitive ethology, developmental psychology and cognitive archeology. Talk of animals having beliefs about conspecifics or infants possessing bodies of knowledge about objects and how they behave should be taken as shorthand for a more complex explanation in terms of the simpler forms of central cognition that we have been discussing" (Bermúdez, 2005, pp. 329-330). If we can explain the results of the current studies in terms of low-level strategies, e.g. template matching, pattern recognition, etc., they don't support the conclusion that nonlinguistic beings possess propositional attitudes. If they don't support that conclusion, however, they neither disconfirm the prediction I attributed to Bermúdez at the beginning of the section nor pose a challenge to his account.

The second substrategy might be described as the nonconceptual content strategy. According to this substrategy, whereas genuine propositional attitudes have conceptual content, the thoughts of nonlinguistic nonhuman animals and prelinguistic human infants have nonconceptual content; rather than actual beliefs, desires, etc., then, animals and infants only have proto-beliefs, proto-desires, etc. As was the case with the first substrategy, if we can explain the results of the above studies in terms of nonconceptual content and proto-propositional attitudes, they don't support the conclusion that nonlinguistic beings possess real, full propositional attitudes. Again, if they don't support this conclusion, they don't disconfirm the prediction I've attributed to Bermúdez or challenge his account.

As it turns out, however, neither of these substrategies seems to be particularly well-equipped to deal with the studies we've been discussing. First, there are at least two difficulties with applying the first substrategy to the studies. The first difficulty is that the researchers in at least some of the studies explicitly reject the suggestion that their results can be explained in terms of low-level processes. For example, both Flombaum and Santos (2005) and Onishi and Baillargeon (2005) consider and reject a range of possible low-level explanations for their findings. The second difficulty is one that Bermúdez (2005) himself raises. As he notes, Hilary Kornblith has argued that, "we have no better perspective than our actual scientific practices for determining the legitimacy of propositional attitude ascriptions" (2005, p. 330). In other words, scientific practice is our best guide to determining whether beings possess propositional attitudes; if our scientific practice seems to support the conclusion that a being possesses such attitudes, then, that gives us good reason to believe that it does.

Second, there are also at least two difficulties with applying the second substrategy to the studies. The first difficulty is that the nature of nonconceptual content is controversial and, according to at least some of the available accounts, it isn't even the type of thing we can use to explain the above results. The second difficulty is that, even if we accept Bermúdez's (2003) own account of nonconceptual content, the second substrategy (arguably) can't be applied to the second set of results I described above. In a separate discussion of nonconceptual content, Bermúdez notes that the "central idea behind the theory of nonconceptual mental content is that some mental states can represent the world even though the bearer of those mental states does not possess the concepts required to specify their content" (2003, para. 1). This seems to suggest that nonconceptual content is to be invoked in cases in which a being has a thought but lacks the concept that applies to the content of that thought, e.g. a being that experienced a mental state that represented a cube but that lacked the concept CUBE would be experiencing a mental state with nonconceptual content. If this is the case, though, the thoughts experienced by the subjects in the second two studies arguably weren't nonconceptual. As Bermúdez (2005) acknowledges in Chapter 7, in addition to indicating that a being can attribute mental states to others, success at the false belief task is typically taken to indicate competence with the concept BELIEF; because it demonstrates an understanding that beliefs can misrepresent, success at the task is taken as evidence of a genuine understanding of the belief concept. The fact that the chimpanzees and human infants in the second two studies passed versions of the false belief task seems to suggest, then, that they possessed the belief concept.[3] This, in turn, suggests that, when they had thoughts about the beliefs of others, i.e. when they attributed beliefs to others, they possessed the concept that applied to the contents of those thoughts, i.e. the concept BELIEF. If they possessed the concept that applied to the contents of their thoughts, though, it doesn't seem appropriate to conclude that those thoughts were nonconceptual; rather, the thoughts seem to have had the conceptual content that's characteristic of real, full propositional attitudes.

In this section, then, I've made two suggestions. First, I've suggested that recent empirical findings give us reason to believe that at least some nonlinguistic beings possess propositional attitudes. This, in turn, suggests that the empirical prediction I attributed to Bermúdez (2005) earlier in the section doesn't hold. Second, I've suggested that, at least at first glance, Bermudez doesn't have any obvious way of accommodating these findings. Of course, it's entirely possible that he might be able to either show that the findings don't have the implications I take them to have, revise one of the above responses to accommodate the findings or introduce a new strategy that diffuses the threat. For at least the moment, though, the studies do seem to disconfirm the

[3] I should note that Hare, Call and Tomasello (2001) don't take their results to definitively demonstrate this.

prediction that nonlinguistic beings will lack propositional attitudes. If this prediction fails to hold, that seems to cast some doubt on Bermúdez's claim that all propositional attitudes are ultimately rooted in natural language. At least pending further argumentation or empirical results, then, we might want to withhold our assent to his specific account of the propositional attitude complex. In other words, until Bermúdez can either show that the above studies don't have the implications I take them to have or that these implications aren't really problematic for his account, we should hesitate to agree that the propositional attitude complex is fundamentally natural language-based.

3 Conclusion

In the preceding sections, I tended to focus on my reservations about Bermúdez's (2005) *Philosophy of Psychology: A Contemporary Introduction*. In general, though, I think that the book serves as an excellent introduction to the subject matter and identifies some interesting avenues for future research. In this conclusion, then, I want to emphasize some of its merits. First, as I see it, there are two primary purposes of an introductory text like the current volume. The first purpose is to offer an accessible account of the subject matter and the second is to encourage further, more in-depth investigation of that subject matter. In my opinion, *Philosophy of Psychology: A Contemporary Introduction* succeeds on both of these counts.

Perhaps most importantly, Bermúdez's (2005) presentation of the issues is consistently very clear. Some of the topics that he covers in the book, e.g. neural nets, could easily have become overly technical. In his discussion of these topics, though, he includes enough layperson-friendly examples and explanations to ensure that they are accessible to non-experts. Even readers who have relatively little prior familiarity with the philosophy of psychology can, therefore, understand and engage with the issues he raises.

Also, the way in which he has structured the book seems, to me, to have some important advantages. As noted in §1, Bermúdez (2005) uses the interface problem to frame his discussion of the four dominant pictures of the mind and the primary issues in contemporary philosophy of psychology. One of the major advantages of this type of framing is that it highlights (and, thereby, clarifies) the connections between both the different pictures of the mind and these pictures and the issues. Another major advantage is that it helps to motivate interest in the issues. By starting from the dialectic between the four dominant pictures of the mind, Bermúdez is able to show why philosophers of psychology should be interested in these particular issues and what exactly is at stake for them in the specific issue debates. In general, then, I think that his choice of a framing device helps to serve both of the purposes I mentioned above. That is, it helps both to render his discussion accessible and to encourage further, more detailed investigation of the issues.

Second, the book offers some interesting suggestions for future investigation of the nature of the mind. Some of these suggestions are implicit. By highlighting the commitments of the different pictures of the mind and the implications of the specific issue debates for these pictures, Bermúdez (2005) hints at some questions that might help us to choose between the pictures and / or resolve the debates. Others are more explicit. As noted in the previous sections, in the last chapter of the book, Bermúdez offers two suggestions about how we should think about the mind. The first suggestion is that we should develop a picture of the mind that includes elements of more than one of the currently dominant pictures and the second is that we should understand the propositional attitude complex in terms of natural language. As I emphasized in the previous section, I have some hesitations about the second of these suggestions. The first suggestion, on the other hand, strikes me as very plausible; whether or not Bermúdez's description of the structure of the mind turns out to be completely accurate, I think his insight that the correct picture might involve elements of multiple paradigms is a valuable one.

On the whole, then, I think that *Philosophy of Psychology: A Contemporary Introduction* makes a significant contribution to the philosophy of psychology literature. Not only does it nicely summarize the current state of the field but it also suggests some interesting directions for future research. I think, therefore, that it could very profitably be read both by students and researchers who were relatively new to the subject matter and by those who were in search of a new perspective on the issues.

References

Bermúdez, J. L. 2003. “Nonconceptual Mental Content.” Web page, [accessed 9/1/2006]. Available at: http://plato.stanford.edu/entries/content-nonconceptual/.

Bermúdez, J. L. 2005. *Philosophy of Psychology: A Contemporary Introduction*. New York and London: Routledge.

Emery, N. J. and Clayton, N. S. 2001. “Effects of Experience and Social Context on Prospective Caching Strategies by Scrub Jays.” *Nature* 414: 443-446.

Flombaum, J. I. and Santos, L. R. 2005. “Rhesus Monkeys Attribute Perceptions to Others.” *Current Biology* 15(5): 447-452.

Fodor, J. A. 1992. “A Theory of the Child’s Theory of Mind.” *Cognition* 44: 283-296.

Goldman, A. I. 2006. *Simulating Minds*. New York: Oxford University Press.

Hare, B. Call, J. and Tomasello, M. 2001. “Do Chimpanzees Know What Conspecifics Know?” *Animal Behaviour* 61: 139-151.

Leslie, A. M. and Polizzi, P. 1998. “Inhibitory Processing in the False Belief Task: Two Conjectures.” *Developmental Science* 1(2): 247-253.

Mynatt, C. R. and Doherty, M. E. 1999. *Understanding Human Behavior*. Boston: Allen and Bacon.

Onishi, K. H. and Baillargeon, R. 2005. “Do 15-Month-Old Infants Understand False Beliefs?” *Science* 308: 255-258.

Zimbardo, P. G. and Weber, A. L. 1997. *Psychology* (Second Edition). New York: Longman.

Commonsense Psychology and the Interface Problem: Reply to Botterill

José Luis Bermúdez

Department of Philosophy
Washington University in St. Louis

George Botterill's challenging and wide-ranging paper raises a number of important issues. Many of these issues cluster around the general theme of how we should think about commonsense psychology and the interface problem. Botterill takes exception; to how I characterize commonsense psychology; to how (in the form of the interface problem) I use commonsense psychology to classify different approaches to the philosophy of psychology; and to the alternative "fifth picture" that I tentatively propose in the final chaper. In this reply I explain why some of Botterill's charges seem to me to fail to miss the mark – while others can actually be put to work in support of my alternative picture.

Let me begin with a general methodological point. Early on in his paper Botterill says the following about my strategy of identifying four different pictures of the mind in terms of four different ways of thinking about how the personal-level explanations of commonsense psychology "interface" with subpersonal levels of explanation lower down the hierarchy of explanation:

> The four pictures are defined in relation to the interface problem. That makes them primarily concerned with the relation between folk or commonsense psychology and how the mind actually operates. While this has been a topic of interest to philosophers, concentrating on that as the central issue may not be the best way of establishing a general paradigm for the cognitive sciences.

Botterill may well be right that anything that has pretensions to be a general paradigm for cognitive science must involve far more than a solution to the interface problem – and he is also right that at least one of the pictures I present (the picture of the autonomous mind) has little or no plausibility as a "general paradigm for the cognitive sciences". But this is not really relevant to what I was trying to do in presenting the four pictures.

I was not proposing the four pictures as four different paradigms for cognitive science. I began (in Chapter 1) with the general idea that there is a hierarchy of levels of explanation, where each level has different *explananda*, different explanatory primitives, and different explanatory generalizations. As we go up through the hierarchy the level of generality increases and we move further away from the details of physical implementation. We might plausibly see molecular biology towards the bottom of the hierarchy and commonsense psychology towards the top. Some of those levels no doubt count as part of cognitive science. Others do not. The task that the existence of this hierarchy poses for the philosophy of psychology is to model how those levels fit together. This is a very different task from establishing a paradigm for the cognitive sciences – just as different models of the unity of science are not *ipso facto* different paradigms for natural science.

So, the four pictures that I put forward are four different pictures of how the hierarchy of explanation fits together. But why, Botterill asks, should we define them in terms of how they think about commonsense psychology? Why pick out commonsense psychology for special treatment? Well, one reason for doing so is that almost everyone else has done it that way. This is a powerful consideration when writing a textbook! But, equally importantly, I think that philosophers have had good reasons for singling out commonsense psychology for special treatment. Commonsense psychological explanations raise a number of *philosophical* problems that either do not appear or are far less prominent at lower levels of explanation. These problems include the causal efficacy of the states and processes posited by commonsense psychology; the role of rationality constraints in commonsense psychological explanation; and the nomological status of the generalizations of commonsense psychology. Some of these problems are explicitly addressed in Chapter 5. Others appear throughout the book. These specific philosophical problems give the interface problem its point, and what I tried to show is that different solutions to them generate different ways of thinking about the hierarchy of explanation. In some cases these different pictures are aligned with different paradigms for cognitive science (as in the neurocomputational picture, for example), but in others they are not (as in the autonomous mind).

Later on in his paper Botterill expresses some skepticism about thinking of commonsense psychology as an autonomous level of explanation at all. He states:

> In my opinion the defect of this approach is that it takes commonsense psychology to be an independent branch of cognition, while describing it in terms which are more appropriate to a *practice*. Moreover, such a practice would be better regarded as the socially communicable outcrop of a more basic mindreading capacity.

He expands on this point by taking issue with my general deflationary proposal (also discussed by Alfredo Paternoster in his comments) to narrow what I term the scope of commonsense psychology.

The aim of the "narrowing strategy" (pursued in Chapter 7) is to identify mechanisms and heuristics that can underwrite social understanding and social coordination without involving the attribution of propositional attitudes – and hence without raising the sort of problems that makes the interface problem so difficult. Botterill objects that I am equating what he calls *mindreading* with propositional attitude psychology, so that it comes out as more or less tautologous that *mindreading* involves deploying representations of propositional attitudes to explain, predict, and control behavior. Mindreading, as Botterill understands it, is more primitive than commonsense psychology. It involves an informational understanding of the mind, of the sort that (some) developmental psychologists have identified in young infants long before they pass the false belief test. To be a mindreader is to treat the behavior of others as guided by the information that they possess about the world – but not necessarily to treat them as guided by beliefs, desires, and other content-bearing propositional attitudes. What is wrong with the "narrowing" strategy, according to Botterill, is that it jumps directly from the sophisticated tools of propositional attitude psychology to completely non-psychological strategies, mechanisms, and heuristics. The strategy leaves out the primitive forms of mind-reading that he (very plausibly) sees as underpinning propositional attitude psychology, and preceding it in the normal course of human development.

Botterill is quite right to highlight the importance of these primitive forms of mind-reading. It seems to me, however, that they can be integrated very easily into the framework I am proposing. In fact, some of them are there already! In Chapter 7 I discuss the primitive mechanisms of emotional sensitivity that play a very important role from the earliest stages of human development. These are perhaps the most basic way in which people can make sense of each other and coordinate their behavior without deploying the complexities of propositional attitude psychology – and they seem to count as forms of mindreading, by Botterill's lights. We can see this in such well-documented developmental phenomena as social referencing (see, e.g., Klinnert et al. 1983). In social referencing infants regulate their own behavior by investigating and taking cues from the emotional reactions of others to a particular situation. An infant who comes across a puzzling or intimidating situation will look towards his mother or other caregiver for guidance and his subsequent behavior is influenced by his perception of her emotional reaction. Social referencing involves treating the caregiver as possessing information about the environment, information that is revealed in her emotional responses.

As Botterill points out, there are many more examples of primitive forms of mindreading. He mentions proto-declarative pointing, gaze-following, and shared attention, among others. I am puzzled as to why he thinks that I would not be quite happy to count these among the simpler mechanisms that serve to narrow the scope of commonsense psychology. For one thing, they quite plainly serve to underpin and facilitate forms of social understanding and social coordination. And, equally plainly, they do not raise the cluster of issues and difficulties that generate the interface problem. Finally, they seem to be precisely the sort of primitive abilities that can best be modeled at the subpersonal level in terms of pattern recognition and template-matching. All in all, the primitive forms of mindreading are very much grist to my mill.

I have written elsewhere about the relation between propositional attitude psychology and primitive forms of mindreading (although not under that label). In *Thinking without Words* I offered an argument that metarepresentation (thinking about thoughts) requires language and hence is only available to language-using creatures. In short, intentional ascent requires semantic ascent. This argument, if sound, means that a very broad class of psychological attributions is unavailable to non-linguistic creatures (including of course prelinguistic human infants). To attribute a belief, for example, to another creature is essentially to view that creature as standing in a particular relation to a thought – the relation of believing the thought to be true. Clearly, therefore, the attribution of a belief requires thinking about a thought. In fact, it has the consequence that propositional attitude psychology is out of the reach of non-linguistic creatures.

Nonetheless, I argued in some detail in *Thinking without Words* (Bermúdez 2003) that there are types of mental state that can be comprehended and attributed by non-linguistic creatures. The analyses there set out to provide a philosophical framework for thinking about types of mindreading that are, of necessity, more primitive than those exploiting propositional attitude psychology. Let me end by sketching out some of the salient details for two such forms of mindreading – the reading of other's desires and the reading of their perceptions (as might be exploited, for example, in joint attention and gaze-following).

We can distinguish two ways of thinking about desire. One can desire a particular thing, or one can desire that a particular state of affairs be the case. I call this the distinction between *goal-desires* and *situation-desires*. Goal-desires are more basic than situation-desires, which fall squarely in the domain of propositional attitude psychology. The contrast is effectively between desire construed as a propositional attitude (in situation-desires, which are attributed via that-clauses picking out the thought that is the object of desire) and the more fundamental goal-desires that are directed not at thoughts but rather at objects or features. There is no reason why non-linguistic creatures should not be able to attribute goal-desires to other agents, since goal-desires are relations between a subject and an object/feature, rather than between a subject and a proposition.

The ability to attribute goal-desires goes hand in hand with a basic understanding of intentional, that is to say goal-directed, behavior – and hence brings with it certain basic mindreading capacities. A purposive action is an action for which a motivating goal-desire can be identified.

Goal-desires cannot be the only mental states that can be identified and attributed by non-linguistic creatures. It is hard to see, for example, how a goal-desire can be attributed to a creature without some evidence of the information that the creature possesses about its environment. At the bare minimum this information will be perceptual. To know what goal-desire might be motivating a creature at a given moment a creature needs to know, first, what end it is pursuing and, second, how it might reasonably expect that end to be realized by its current behavior. Both of these require knowing to which features of its environment the creature is perceptually sensitive. If, therefore, a non-linguistic creature is to be able to attribute goal-desires to a fellow creature it must be able to formulate hypotheses about what that creature is perceiving.

Here too we can distinguish two ways of thinking about seeing by following Dretske in the distinction between *simple seeing* and *epistemic seeing*. According to Dretske, what we see in simple seeing (or what he calls non-epistemic seeing) "is a function solely of what there is to see and what, given our visual apparatus and the conditions in which we employ it, we are capable of visually differentiating" (Dretske 1969: 76). In contrast, epistemic seeing involves standing in a relation to a proposition (a thought). Epistemic seeing involves seeing *that* something is the case – and, because of this, the attribution of epistemic seeings is an exercise in propositional attitude psychology.

Even if one thinks (as I do) that non-linguistic creatures are not capable of understanding epistemic seeing, since this involves thinking about the perceiver's relation to a thought, this is perfectly compatible with non-linguistic creatures being capable of thinking about the direct perceptual relations in which other creatures stand to objects. This makes available a further form of mindreading and allows non-linguistic creatures to engage in a primitive form of psychological explanation. A creature that knows what a conspecific or predator desires and has some sense of its perceptual sensitivity to the environmental layout (as well as an understanding of its motor capabilities) can expect to be able to predict its behavior with some success.

This restrictive interpretation of the "mind-reading" abilities of some non-linguistic creatures is compatible with much recent research into the extent to which non-human primates can properly be described as possessing a "theory of mind". There are well-documented examples of primate behavior that some prominent students of animal behavior have thought can only be interpreted as examples of interpersonal deception (see, e.g., Premack and Woodruff 1978 and the papers in Byrne and Whiten 1995). But the consensus opinion among primatologists is that a more parsimonious interpretation of these behaviors is to be preferred (see, e.g., Povinelli 1996 and Hauser 2000).

It should be clear, then, that I am very sympathetic to Botterill's emphasis on forms of mindreading that are more primitive than propositional attitude psychology. Understanding these primitive forms of mindreading is essential to understanding why the interface problem may not be as all-encompassing as philosophers have taken it to be. I remain convinced, though, that the interface problem is the best place to start in thinking about the philosophy of psychology.

References

Bermúdez, J. L. 2003. *Thinking without Words*. New York. Oxford University Press.

Byrne, R. W and Whiten, A. (Eds.) 1988. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*. Oxford. Blackwell.

Hauser, M. S. 2000. *Wild Mind: What Animals Really Think*. London. Penguin Books.

Dretske, F. 1969. *Seeing and Knowing*. London. Routledge.

Klinnert, M. D., Campos, J. J., Sorce, J. F. Emde, R. N. Svejda, M. 1983. "Emotions as Behaviour regulators: Social Referencing in Infancy". In R. Plutchik and H. Kellerman (Eds.) 1983. *Emotion: Theory, Research, Experience*. Boston. Academic Press.

Povinelli, D. J. 1996. "Chimpanzee Theory of Mind". In P. Carruthers and P. K. Smith (Eds.), *Theories of Theories of Mind*. Cambridge. Cambridge University Press.

Premack, D. and Woodruff, G. 1978. "Does the chimpanzee have a theory of mind?" *Behavioural and Brain Sciences* 1: 515-526.

Counterfactuals and Token Identity: Reply to Lowe

José Luis Bermúdez

Department of Philosophy
Washington University in St. Louis

Jonathan Lowe's very interesting comments contain an ingenious argument against the thesis that mental events are token identical to physical events. He takes the example of a mental act of choice, which he plausibly describes as appearing to introspection as a single and unitary occurrence initiating a particular bodily action. The thesis that this mental act of choice can be identified with any neural event is, he argues, decisively defeated by considerations from standard ways of thinking about counterfactual conditionals. He proposes this in support of an extreme version of what I called the picture of the autonomous mind – a version on which the explanatory incommensurability between the personal and subpersonal levels reflects an ontologically dualist distinction between persons and their bodies. In this brief reply I argue that Lowe's argument commits a subtle, but nonetheless fallacious, equivocation, and that an identity theorist has the resources to accommodate the phenomena to which he draws attention.

Let me begin by recapitulating Lowe's argument. If the mental act of deciding to raise my arm is to be identical to anything, it could only be to a highly complex and multiply ramified neural event. *Prima facie*, Lowe suggests, individual mental acts have rather different causal profiles from complex neural events. Consider the act of deliberatively deciding to raise my arm, and any neural event that might be a candidate for being identical to that act: "The act of choice seems to be unitary and to have, all by itself, an 'initiating' role, whereas the neural events seem to be thoroughly *disunified* and merely to contribute in different ways to a host of different ongoing causal chains, many of which lead independently of one another to the eventual arm movement".

The prima facie appearance of non-identity can, Lowe claims, be buttressed by an argument that explanations invoking the mental act and the complex neural event have very different counterfactual implications. Let 'D' and 'N' name the mental act and neural event respectively, and let B stand for the bodily movement of my raising my arm. Any identity theorist must accept the following counterfactual: If N had not occurred, then B would not have occurred. On standard models of the semantics of counterfactuals this means that, in the nearest possible world in which N does not occur, B does not occur. So, Lowe asks, what sort of event occurs instead of B in the nearest possible world in which N does not occur?

The nearest possible world in which N does not occur is, he argues, a world in which some other, very similar neural event, N*, occurs. N* overlaps significantly with N. It is N, plus or minus a couple of neurons. Given this, Lowe argues that the nearest N*-world to this one is a world in which N* causes a bodily event B* that is very similar to B – certainly similar enough to count as an arm-raising. So, the nearest possible world in which N does not occur is a possible world in which I still raise my arm. Yet, if my decision to raise my arm really is to count as a cause of my arm rising then, Lowe maintains, the nearest possible world in which I do not decide to raise my arm is a world in which my arm does not go up. So, given that the nearest not-N world is a B-world, while the nearest not-D world is a not-B world, Lowe concludes that D cannot be identical to N.

The problem with Lowe's argument emerges when we ask about the kinds under which each of the events he discusses fall. Let me define the N-kind as the kind under which fall those events that consist of almost exactly the same neural events as N and that could easily have occurred in place of N. Plainly, Lowe's argument depends upon the N-kind having at least one member, namely N*. It also depends upon the B-kind having at least one member, where the B-kind is defined as those bodily movements that are almost exactly the same as B and that could easily have occurred in place of B. Suppose now that we define the D-kind to include those decisions that are almost exactly the same as D and that could easily have occurred instead of D. Lowe's argument is driven by the assumption that the D-kind is empty: "It is quite incredible to suppose that if the agent had not made *that* very decision, D, then he or she would have made another decision virtually indistinguishable from D – in other words, *another* decision to raise the arm in the same, or virtually the same, way."

I note first that the assumption that the D-kind is empty seems to beg the question against the identity theorist. A theorist who thinks both that D = N and that the N-kind is non-empty, will of course conclude that the D-type is equally non-empty. After all, if D = N then anything that occurs (or could occur) in place of N equally occurs (or could occur) in place of D. But it is most unclear where the burden of the argument lies, since Lowe would most likely reply that one theorist's modus ponens is another's modus tollens. If it is as obvious as Lowe thinks that the D-kind is empty, then running the same line of reasoning in the opposite direction seems clearly to show that D cannot be identical to N.

But *is* it as obvious as Lowe thinks that the D-kind is empty? Here I have my doubts. Let us look again at what he says:

> It is quite incredible to suppose that if the agent had not made *that* very decision, D, then he or she would have made another decision virtually indistinguishable from D – in other words, *another* decision to raise the arm in the same, or virtually the same, way. On the contrary, if the agent had not made *that* decision, then he or she would either have made a quite different decision or else no decision at all. Either way – assuming that there is nothing defective in the agent's nervous system – the arm *would not* have risen almost exactly as it did.

Let me distinguish two ways of thinking about the decision to raise my arm in order to catch a lecturer's attention. What is relevantly different in each case is the contrast class of things that I might otherwise have done. We might think of it, first, as a decision to raise my arm *as opposed, say, to attracting the lecturer's attention in some other way*. Let me call this a coarse-grained decision. We can also think of the decision in a fine-grained way, where what I decide to do is to raise my arm along a particular trajectory φ, *as opposed, say, to raising my arm along trajectory* ψ.

Now, it is plain that, had I not made the coarse-grained decision to raise my arm, then no event similar to B would have occurred. This does not support Lowe's argument, however, because no identity theorist is likely to claim that the coarse-grained decision is identical to some complex neural event N. The identity claim is more naturally interpreted as holding between N and a fine-grained decision, so the emptiness of the D-when the decision is interpreted in the coarse-grained manner is not problematic. And, contrary to Lowe, it seems very plausible that the nearest possible world in which I do not make the fine-grained decision to raise my arm along trajectory φ_i is in fact a world where I make another fine-grained decision to raise my arm along the rather similar trajectory φ_j. When we think about D in the fine-grained sense, therefore, there are good reasons for thinking that the D-kind is non-empty.

The distinction between coarse- and fine-grained decisions allows the identity theorist to do justice to the phenomena that Lowe tries to capture with his distinction between intentional causation and bodily causation. According to Lowe, intentional causation (the sort of causation that can be effected by decisions, and other mental events) is fact causation, whereas bodily causation is event causation. My decision causally explains the fact that a bodily event of a certain type occurred, whereas the occurrence of the particular bodily event is causally explained by some complicated neural event. On the alternative I am proposing, what Lowe characterizes as a relation of fact causation holding between a non-physical decision and a fact should be viewed instead as a relation between a coarse-grained decision and a set of bodily movements.

Of course, this only counts as an advance if we have a clear understanding of the relation between coarse-grained decisions, fine-grained decisions, and bodily movements. My proposal is the obvious one. All decisions are, strictly speaking, fine-grained (so that it is only fine-grained decisions to which the identity theory applies directly). When we think of decisions in the coarse-grained way what we are really doing is abstracting away from many of the crucial details of particular fine-grained decisions. We are focusing on what is consciously accessible (the goal and/or target of the movement, for example), rather than the implementational details of precisely calculated limb movements and muscle contractions that take place below the threshold of awareness. I am less inclined than Lowe to give weight to the introspective "datum" that decisions are unitary phenomena. It seems much more likely to me that decisions are highly complex processes that incorporate highly detailed movement planning and complex forms of information-processing. We are at best only very partially aware of all that goes on when we decide to do something.

The question that Lowe thinks can only be answered by intentional causation (the question of why a bodily event of *this* type occurs) is really the question of why all the bodily events that could have occurred share a certain high-level abstract description. I imagine that in most cases there is a relatively straightforward explanation in terms of preceding events and the general context of the action. Among the factors feeding into a particular fine-grained decision, such as the decision to raise my arm, are both psychological factors (the desire to attract the lecturer's attention, for example) and physical factors (such as the starting-position of my arm, the tiredness of my shoulder muscles, the existence of obstacles blocking certain possible trajectories, and so on). It is, one might plausibly conjecture, constancies in the first group and variations in the second that account for why there are many different bodily events that could have occurred, all of which share a certain high-level description.

So, to return to Lowe's argument against the identity theory, I claim that it rests upon an equivocation in the notion of a decision. Lowe denies the following counterfactual: Had decision D not occurred, then some bodily event B* very similar to B would have occurred. This denial is legimitate only if D is understood in the coarse-grained sense. But, strictly speaking, there are no coarse-grained decisions. Coarse-grained decisions are simply abstract characterizations of fine-grained decisions and, when we take 'D' in the counterfactual to denote a fine-grained decision, the counterfactual comes out as true. This is exactly what one would expect, if (as the identity theorist claims) the fine-grained decision is identical to some complex neural event N and (as both Lowe and the identity theorist accept) it is true that, had N not occurred, then some bodily event very similar to B would have occurred.

The Interface Problem and the Scope of Commonsense Psychology: Reply to Paternoster

José Luis Bermúdez

Department of Philosophy
Washington University in St. Louis

Alfredo Paternoster raises a very interesting objection to one aspect of the "fifth picture" of the mind sketched out in the final chapter of *Philosophy of Psychology: A Contemporary Introduction*. In this short reply I recapitulate the proposal that Peternoster criticizes; summarize his objection; and finally explain why in the last analysis I remain (relatively) unmoved.

The book is structured around four different ways of responding to what I term the *interface problem*. Each of these four pictures of the mind offers a different account of how commonsense psychological explanations of behavior interface with the types of explanation and models to be found in the various cognitive and behavioral sciences. I suggest in the final chapter that, although each of these accounts seems to work well for some aspects of cognition, none of them seems a plausible candidate for a comprehensive and complete account of how the mind works. The alternative tentatively proposed in the final chapter for making progress beyond the four pictures contains a number of distinct proposals. Some of these are intended in a deflationary spirit. In particular, I suggest that philosophers of mind and philosophers of psychology may have attached too much importance to the role of propositional attitude psychology in making sense of other people's behavior and making possible coordinated behavior.

Even the most strident critics of the validity and cogency of commonsense psychology (such as the small handful of *eliminativists* about commonsense psychology) hold that it does in fact play a central role in how we make sense of and interact with each other. One of the principal themes of *Philosophy of Psychology* is that this near–unanimity may be misplaced. Commonsense psychology may play far less significant a role in social understanding and social coordination than standardly assumed. If we follow standard practice and define commonsense psychology as invoking propositional attitudes in the explanation and prediction of behavior, many of our social interactions may be governed by procedures and heuristics that do not engage what some philosophers call the "propositional attitude system". The examples I discussed include simple, non-psychological heuristics such as TIT-FOR-TAT; scripts and routines that allow us to identify and exploit patterns in other people's behavior; and primitive mechanisms that "attune" us to the moods and emotional states of others, as well as to the "affordances" of objects and situations. These mechanisms and heuristics involve relatively simple mechanisms of template-matching and pattern recognition of the sort that artificial neural networks model so well.

This way of thinking about the personal-level dimension of social interaction and social coordination has potential implications for how we think about the architecture of cognition. Two of the four pictures I discuss (the functional mind and the representational mind) are naturally allied with what Paternoster aptly calls the *sandwich model* of the mind, where a central and non-modular propositional attitude system is sandwiched between modular input and output systems. In contrast, I suggest that it may make more sense to replace the standard distinction between central and peripheral processing with a three-way picture on which we can identify two fundamentally different forms of central cognition, in addition to the peripheral modules responsible for processing sensory input. Personal level cognition can involve either the complex processes and mechanisms defined over the propositional attitudes or much simpler mechanisms of template-matching and pattern recognition. I suggest that there are two fundamentally different personal-level routes to action, one engaging the propositional attitudes and the other engaging evolutionarily more primitive mechanisms that are faster and more specialized. We can think of the propositional attitude system as superimposed upon a complex network of pathways leading from peripheral input modules to peripheral output modules. Each pathway leads from input modules to output modules without engaging the propositional attitude system.

One of the advantages I claimed for this alternative picture is that it makes the interface problem far less pressing by "downsizing" the role of commonsense psychology both in personal-level cognition and in subpersonal cognitive architecture. It is to this "deflationary" move that Paternoster takes exception. He presents his concern by focusing on the picture of the mind that most strongly emphasizes the differences between personal-level commonsense psychological explanations and subpersonal information-processing explanation. According to proponents of the picture of the autonomous mind, there is a radical incommensurability between the two modes of explanation. Autonomy theorists are driven to postulate the radical incommensurability, according to Paternoster, because of the following thesis:

> **Failure of supervenience:** Mental (= personal-level) properties do not supervene either on computational properties or on other natural properties.

In effect, Paternoster presents me with a dilemma. The first horn of the dilemma comes with his claim that no theorist attracted to the picture of the autonomous mind by the failure of supervenience thesis is likely to be satisfied by my deflationary proposal. The second horn is that to reject the failure of supervenience thesis is essentially to deny that there really is an interface problem. So, in essence, either the interface problem remains untouched by my proposal, or it was never really a problem at all.

The first horn of the dilemma is the more interesting of the two. (After all, in general terms, it cannot be the case that anyone who thinks that they've solved a problem is committed to there not having been a problem in the first place!) Here is how Paternoster states the difficulty:

> The interface problem is the problem of linking personal-level predicates, hopefully *all* personal-level predicates, to subpersonal-level descriptions. The idea of substituting, for instance, a heuristic not based on belief-desire attributions (assuming, for the sake of the argument, that these heuristics do not involve psychological reasoning at all) does not affect this problem, since not only beliefs and desires, but also mental images, perceptions, and emotions have some role to play in determining behavior, and these are all personal states.

He illustrates the problem with a specific example. We are to assume that a commonsense psychological explanation (that someone asked for a steak, say, because they *believed* that the person in front of them was a waiter and they *desired* to communicate their order) has been replaced with a more deflationary explanation (that she ordered the steak as the appropriate "move" in the "restaurant script"). All that has happened according to Paternoster is the the problem has been shifted sideways:

> We can say something like the following: X has performed the action A, say, asking for a steak, because he has performed a template-matching T (ie he has performed something allowing him to recognize the person in front of him as a waiter). T, in turn, is realized by such and such an algorithm. The problem with this picture is that running a template matching is not a predicate that can be attributed to an *agent*. The correct description of the situation is rather the following: X performed the action A because he was in a recognitional state R. R is realized by a template matching T. *We* recognize, whereas our *minds/brains* perform template-matching. Even if the recognitional state R is not a *belief*

> state – which I am absolutely prepared to concede – still, there are two different levels in question: one personal and the other subpersonal, and the problem of clarifying the relation between the two levels is still there.

Paternoster's point would be very compelling if the interface problem were indeed motivated in the way he suggests it is. If the force of the interface problem rests upon the failure of supervenience thesis, then replacing one personal-level state for another is not going to resolve it. The problem of how personal level explanations relate to subpersonal level explanations, given the failure of supervenience between personal-level states and subpersonal-level states, still stands.

The problem, though, is that the interface problem is not driven by the failure of supervenience thesis. Recall that Paternoster is trying to do justice to the complaints of incommensurability between the personal and subpersonal levels of explanation raised by autonomy theorists. In Chapter 5 of *Philosophy of Psychology* I discuss two powerful articulations of the charge of incommensurability. One is due to Donald Davidson, whose central claim is that there are not (and cannot be) psychophysical. Personal level explanation contains only non-strict generalizations, whereas strict laws are at least in principle available at the subpersonal level. Davidson does not accept the failure of supervenience thesis. Quite the contrary, in fact. Davidson is quite happy to accept that psychological properties supervene on physical properties. In 'Mental events' he explicitly adopts a version of what is now generally known as weak supervenience (to the effect that, necessarily, any two physically indiscernible individuals are psychologically indiscernible). The same holds for Daniel Dennett. The version of the incommensurability thesis that Dennett proposes in real patterns is perfectly compatible with some version of supervenience (in fact, it involves a supervenience thesis in all but name).

What drives the versions of the incommensurability thesis proposed by Davidson and Dennett is in fact considerations from the norms that govern personal-level explanation. Davidson's argument for the anomalism of the mental, as is well known, is based on considerations of the holism of the mental and the open-ended nature of the norms of rationality. What makes personal-level explanation so different from subpersonal explanation is that it involves content-bearing states that not only cause but also rationalize the behavior that they explain. The precise details of his argument for the impossibility of psychophysical laws are notoriously elusive, but what is plain is that the argument exploits distinctive features, not of explanations that exploit personal-level states in general, but rather of explanations that invoke propositional attitudes.

Something similar is true of Dennett. Dennett's basic point in 'Real patterns' is that personal-level explanation exploits patterns in behavior that are invisible when we move to a lower level of explanation. These patterns are real in the sense that they yield explanatory and predictive leverage. Dennett claims that they track genuine causal powers. But they cannot be mapped onto patterns identifiable at lower levels of explanation. They are self-standing and autonomous. As with Davidson, the autonomy of these personal-level patterns is driven by the role that considerations of rationality play in personal-level explanation. And the relevant explanations, once again, are propositional attitude explanations.

So, if what drives the incommensurability thesis (and, by extension, the perceived significance of the interface problem) is the role that norms of rationality play in propositional attitude explanations, then one might expect that the force of the thesis and the significance of the problem would indeed be diminished if ways are found to replace propositional attitude explanations with explanations that do not involve propositional attitudes. I take it that explanations invoking the simple heuristics and mechanisms that I identify do not exploit or depend upon principles of rationality in ways that create the sort of difficulties that Davidson and Dennett highlight. Of course, as Paternoster points out, some account still needs to be given of how these heuristics and mechanisms are implemented at the personal level. To that extent the interface problem still stands. But, I suggest, Paternoster is wrong to suggest that this new interface problem is just as intractable as the old one.

Do non-linguistic creatures possess second-order propositional attitudes? Reply to Shanton

José Luis Bermúdez

Department of Philosophy
Washington University in St. Louis

Karen Shanton's well-informed and insightful commentary raises the question of whether the account of the high-level propositional attitude complex that I propose in the final chapter of *Philosophy of Psychology: A Contemporary Introduction* is compatible with recent studies of nonhuman primates and prelinguistic infants. Since my view is that entertaining second-order propositional attitudes (propositional attitudes about another thinker's propositional attitudes, such as the belief that she desires something, or the desire that she believe something) depends upon taking attitudes to natural language sentences, it plainly entails that only language-users can have second-order propositional attitudes. However, as Shanton points out, a number of researchers have made strong claims about the metarepresentational capacities of nonlinguistic creatures. For example, Flombaum and Santos offer the following description of their experiments on free-ranging monkeys, which were designed to show that monkeys use information about where two human "competitors" are each looking in order to choose which one to rob of a grape.

> Rhesus monkeys correctly use information about what a competitor can and cannot see in order to retrieve a contested piece of food. Because the monkeys in these experiments must selectively avoid the experimenter who could potentially see the contested food item, it is difficult to interpret these results in terms of a simple mechanism for responding to the gaze of another individual without representing that individual's perceptions. The animals in these studies needed to first represent what the two competitors could and could not see, and then to make a choice, based on this knowledge, to approach the experimenter who was not visually aware. Consequently, beyond demonstrating that rhesus monkeys are *sensitive* to eye-gaze direction, these experiments constitute the first evidence that a non-ape species spontaneously reasons about another individual's *visual* perception. (Flombaum and Santos 2005, 449)

Onishi and Baillargeon make an even stronger claim about the metarepresentational capacities of 15-month infants, whom they describe as able to pass a version of the false belief task.

> These results suggest that 15-month-old infants already possess (at least in a rudimentary and implicit form) a representational theory of mind. They realize that others act on the basis of their beliefs and that these beliefs are representations that may or may not mirror reality. (Onishi and Baillargeon 2005, 257)

If these (and the other studies that Shanton cites) are indeed to be interpreted in the way their authors propose, then they certainly seem to present problems for my view. However, as I try to bring out in this short note, an alternative interpretation is available.

According to Shanton, there are two ways a defender of my position might respond to experimental evidence of this type. The first strategy is to try to argue that the subjects in the studies really do have language. This is plainly hopeless. More promising, she thinks, is the strategy of denying that the subjects possess propositional attitudes at all. This can be pursued either by claiming that the experimental behaviors can all be explained in terms of low-level capacities for template-matching and pattern recognition, or by claiming that, in contrast to the *conceptual* thoughts of language-using creatures, the thoughts of nonlinguistic creatures only have *nonconceptual content*. Shanton finds neither sub-strategy convincing.

I am not entirely persuaded by her arguments against the two sub-strategies. But for present purposes I am happy to put them to one side, since it is most unclear to me that I need adopt either of them to avoid coming into conflict with the experimental data. None of the views that I present in *Philosophy of Psychology: A Contemporary Introduction* (or in my earlier book, *Thinking without Words* (Bermúdez 2003), which aims to provide a conceptual framework for interpreting the thoughts of nonlinguistic creatures) requires me to deny that non-linguistic creatures can have propositional attitudes. Quite the contrary. In both books the claim is that language is required for propositional attitudes of a particular type – namely, second order propositional attitudes. In *Philosophy of Psychology* I propose that certain types of thinking are only possible as a function of the cerebral rewiring that comes with the emergence of language. These are, first, thinking that requires integrating information in different representational formats, and, second, thinking that takes as its object thoughts (as opposed to objects and properties in the distal environment). This leaves open, of course, the possibility that non-linguistic creatures may have thoughts about objects and properties in the distal environment that do not involve integrating different representational formats but that are more cognitively sophisticated than low-level capacities for template-matching and pattern recognition. In *Thinking without Words* I showed how this possibility is realized in many different types of non-linguistic creature. Non-linguistic creatures can certainly have beliefs and desires. They just can't have beliefs and desires with metarepresentational contents – viz. beliefs and desires

Let me briefly recap the central claim from Ch. 8 of *Thinking without Words*. By a higher-order thought I mean a thought that takes another thought as its object. Thoughts about another's mental states count as higher-order thoughts, for example, as does reflection on one's own mental states. Quine once described *semantic ascent* as "the shift from talking in certain terms to talking about them" (Quine 1960, 271). By analogy we can characterize *intentional ascent* as the shift from thinking in certain ways to thinking about those ways of thinking. My claim, in effect, is that intentional ascent requires semantic ascent – that we can only think about thoughts through thinking about words. The argument for this hinges on the fact that thoughts must be represented at the personal level if they are to be the objects of higher-order thoughts. There are all sorts of things going on below the threshold of consciousness when we think (perhaps thinking involves manipulating sentences in a subpersonal language of thought, for example). But these subpersonal events are not what we think about when we think about our own thoughts. When we think about thoughts (in attributing a thought to another subject, for example) we need to do justice to the structure and to the inferential role of the thought and this requires representing it in a vehicle that makes its structure and inferential role perspicuous. The only possible candidate, so I argued, is a public language sentence.

The real issue, therefore, is how exactly we should understand the content of the thoughts attributed to the experimental subjects in studies such as those to which Shanton draws attention. Plainly, it is not open to me to attribute to them thoughts with metarepresentational contents. So, can they be interpreted differently? Are there ways in which non-linguistic creatures can think about the mental states of others without attributing to them full-fledged thoughts.

The final chapter of *Thinking without Words* contains a number of models for thinking about how non-linguistic creatures could engage in primitive forms of psychological explanation that do not require or involve attributing thoughts. Particularly relevant here is the distinction that I made, following Fred Dretske's distinction between *simple seeing* and *epistemic seeing*, between two ways in which one creature might represent the perceptual states of another. According to Dretske, what we see in simple seeing "is a function solely of what there is to see and what, given our visual apparatus and the conditions in which we employ it, we are capable of visually differentiating" (Dretske 1969, 76). In contrast, epistemic seeing involves standing in a relation to a proposition (a thought). Epistemic seeing involves seeing *that* something is the case. By extension we can distinguish between an understanding of another subject's perceptual states that involves attributing to that subject a relation to a thought, on the one hand, and one that involves identifying those features of the environment that the subject can discriminate. Attributions of the first type are *opaque*, whereas attributions of the second type are *transparent*.

The argument from intentional ascent shows that non-linguistic creatures are not capable of understanding epistemic seeing, since this involves thinking about the perceiver's relation to a thought. So, for example, in the Flombaum and Santos experiments, the rhesus monkeys cannot be attributed the thought that the experimenter can see that the food item is within reach. This would require the monkey to represent the thought that the food item is within reach in order to attribute it to the experimenter (which is, of course, a far more sophisticated cognitive achievement than simply representing the food item as being within reach). But there is no obstacle to attributing to them thoughts about the direct perceptual relations in which other creatures stand to objects – thoughts that track another creature's perceptual sensitivity to the layout of its environment. And there is nothing, I submit, about the data that Flombaum and Santos provide that suggests that anything more than this is going on. It is probably true that the monkeys are doing more than simply tracking eye-direction, but that does not mean that they are attributing thoughts to the experimenters.

In my view, then, the best "deflationary" strategy to take with respect to the experimental results that Shanton cites is not to try to interpret them in completely non-psychological terms, but rather to interpret them as involving primitive forms of psychological explanation that fall short of full-fledged metarepresentation. No doubt it will take considerable detailed work remains to assimilate all the experiments on non-linguistic social cognition within this framework. But I am confident both that this can be done, and that doing it will defuse the tension that Shanton identifies in her comments.

References

Bermudez, J. L. 2003. *Thinking without Words*. Oxford: Oxford University Press.

Dretske, F. 1969. *Seeing and Knowing*. Chicago: University of Chicago Press.

Flombaum, J.I. and Santos, L.R. 2005. "Rhesus Monkeys Attribute Perceptions to Others." *Current Biology* 15(5): 447-452.

Onishi, K.H. and Baillargeon, R. 2005. "Do 15-Month-Old Infants Understand False Beliefs?" *Science* 308: 255-258.

Quine, W. V. O. 1960. *Word and Object.* Cambridge MA: Harvard University Press.