Why can mental states be reduced to neuroscientific explanations but not eliminated?

Abstract

The topic of reduction of mental states to neural explanations has taken in recent years a naturalistic turn, moving from traditional metaphysical arguments and insights, to an interest in how scientific practice itself deals with the question. Two of the more important trends within this new philosophical perspective on psychoneural reduction are the New Mechanistic view and Ruthless Reductionism. The present paper tries to bring a new naturalistic perspective into the discussion that embraces some core ideas of the other two. Central to this new proposal is the suggestion of a model-theoretic framework reflecting the structure and pragmatic constraints underlying the deployment of neuroscientific explanations of behaviours. By means of this framework, it is argued that if neuroscientific explanations are to be relevant and robust, two conditions have to be met: 1- mental states have to be correlated/reduced to specific causal-neural mechanisms and 2- such mental states hold an important methodological-heuristic role even when a causal-neural explanation is in place. This latter consequence directly challenges one of Bickle's ruthless reductionism explicit aims: to show that cellular-molecular explanations render previous mental/psychological explanations *otiose*. This clash between my approach and Bickle's is illustrated and developed throughout the paper.

1-Introduction

In several places, John Bickle claims that current neuroscientific practice provides actual cellular/molecular reductions of certain mental states. He cites the case study of memory consolidation switch as an example where recent findings suggest that this mental state/process can be reduced to the molecular cAMP-PKA-CREB Pathway. Taking this example, Bickle 'waves the eleminativist flag' by claiming that psychological explanations lose their pertinence (or, as he says, "become otiose") once a cellular/molecular explanation replaces them. Some authors have criticised this eleminitavist claim on the basis that neuroscientific explanations consist of mechanistic explanations spanning several levels of organisation/explanation (Craver, 2007; Wright, 2007).

My target in the present paper is *not* Bickle's specific claim for a privileged level of explanation but his more general statement that *within* the frame of current neuroscientific practice, mental states are reducible to neuronal explanations (here, for my purposes, regardless of the level employed by those explanations) and his additional remark that they are also eliminated or rendered otiose once the reduction is achieved. I will present a novel (naturalistic) philosophical approach to the nature of neuroscientific explanation using a semantic-model framework. I will try to show that in order to maintain the power and scope of those explanations one is led to the conclusion that Bickle is right concerning *reductionism* but wrong with respect to *elimination*. I will take Bickle's own example and claim that, even if a reductive explanation of 'memory consolidation switch' is disposable, we *cannot* eschew reductively its causal/functional integrity, i.e. the explanatory/causal context that defines the mental concept/process 'memory consolidation switch' in the first place.

2-Intentional/Cognitive Models in Current Neuroscientific Practice

My basic aim in the present paper is to clarify and understand the role of mental/psychological states in contemporary neuroscientific practice. In order to achieve that goal I will devise a synoptic and general view of how current neuroscience deals with explanatory contexts in which mental vocabulary is used (namely in the explanation of behaviours). It should be stressed that my stance on neuroscientific explanation is a pragmatic and naturalistic one, as it is for Bickle himself. In fact, Bickle is very clear on this methodological point (which he calls New Wave Metascience):

> "The job of new wave metascience is simply to illuminate concepts like reduction as these imbue actual scientific practice. To what end? *Not* to achieve some better way of addressing reformulated 'external' questions about the existence and nature of 'theory-independent ontology'. Rather, because a reasonable explanatory goal is to understand practices 'internal' to important current scientific endeavors and the scope of their potential application and development."
>
> (2003, p.32).

So, Bickle's explicit purpose is to describe results *within* neuroscientific practice, and to extrapolate from them. I follow this naturalistic/pragmatic path in my presentation and discussion of a model-based framework of neuroscientific practice (what I call CN Models) and subsequent reflection on the status of mental states. In doing so, I am following some proposals made by Richard Boyd (1999), in that my approach to mental states conceives them as a vocabulary deployed within a *disciplinary matrix* which, in the present case, includes knowledge and practices from neuroscience, which itself is framed within a larger disciplinary spectrum, including almost all biology (including information from evolutionary theory).

2.1- *Structure and Scope of CN Models*

I will adopt here a model/semantic approach to neuroscientific practice (specifically, in the present context, that part of neuroscientific practice dealing with the explanation of behaviours, hence I will call *this* kind of activity c*ognitive neuroscience* – or CN for short). The semantic view I adopt here has some connections with Ronald Giere's approach to physics (Giere, 1988, 1999). The semantic approach to science endorsed by Giere focuses on scientific activity as a *practice* and how this practice is achieved and carried out by scientists as human cognisers. Likewise other model/semantic approaches to science it is an explicit reaction against the Deductive-Nomological conception. In particular, Giere is quite sensitive to the way mainstream physics is academically communicated through textbooks (Giere, 1988, chap. 3). According to him, the main media of information displayed by those books are idealised physical entities and systems that do not exist in the world (e.g. frictionless motion). So, Giere's proposal is to consider those objects and systems as abstract entities that constitute theoretical models of Classical Mechanics. These models have the status of abstract cognitive entities capable of being represented in several ways: equations, linguistic descriptions, graphic representations deployed by physicists. As Giere states:

> "[T]hey function as 'representation' in one of the more general senses now current in cognitive psychology. Theoretical models are the means by which scientists represent the world – both to themselves and for others."
>
> (1988, p.80).

In this sense, models are *interpreted* intentional objects and not just structures as sometimes proposed.[1]

Just like models of classical mechanics, models of cognitive neuroscience (CN models) are intentional objects, cognitive representations used by scientists in several ways (here too, they can be deployed verbally, graphically or in other ways). More

[1] Current literature on the model-theoretical approach to science is vast and diverse. The word 'model' can mean very different things depending on the approach adopted. Sometimes the term 'model' is used to refer to entities or practices that *help* in the process of constructing explanation (as in the case of 'scale models'). Differently from this perspective my usage of the term 'CN models' refers to the *explanations* themselves. On this view, animal experimental models are part of the evidence in constructing CN models conceived *qua* explanations and so are not covered by my intended usage of the term 'CN model'.

precisely, CN models state general explanations of behavioural phenomena. These models are achieved by inductive abstraction from previous neuroscientific empirical results. Usually, these intentional models are generalisations from animal experimental findings. For instance, results from protocol experiments of spatial memory in mice serve as a *basis* for constructing the explanatory content of a CN model for spatial memory covering all animals that display this mental state, assuming the conservation across species of the relevant structures (in this case, homologues of the CA1 area of the hippocampus in vertebrates). I will address this important topic in more detail later on. For now, let's consider briefly the general structure of these models. Structurally, a CN model (*M*cn) is an ordered triple:

$$M\text{cn}=<\text{B},f,\text{M}>$$

where B is a target behaviour (if we take again the example of spatial memory, it could state something like: "the optimisation in spatial navigation of animals in their environment"), *f* is a mental state explaining B (e.g. spatial memory) and M is a description of a neuronal mechanism explaining *f* (e.g. the mechanism of LTP in hippocampal place cells). Extensionally, *M*cn range over a domain **D** of terrestrial multicellular animals.

So, CN models state *general* neuroscientific explanations by abstraction from *particular* cases (or limited sets of particular cases), i.e. these models are constructed inductively satisfying the normative constraint of *maximising projectability*. The satisfaction of this constraint manifests itself in two important features concerning the nature of CN models: 1- a realist stance regarding mental states and 2- the adoption of a hierarchical taxonomy of those mental states. Let me try to clarify these two important features separately.

2.1.1- *Realism of mental states*

Mental states *f* are real in the sense that they correspond to particular neural mechanisms described by M conceived as real causal features in the world, according to our best data on the subject (i.e., neuroscience). In practice, neuroscientists conceive mental states as (supposedly) realised by certain neural/componential mechanisms that they try to discover. The following extract from Squire, on the nature of declarative memory and its relation to the hippocampus, can be taken as a typical exemplification of this assumption:

> "[T]he terms 'explicit' memory and 'declarative' memory, when one considers the properties that have been associated with each, describe a *biologically real* component of memory that depends on particular structures and connections in the brain."
>
> (1992, p.205, emphasis added).

Neuroscientists constantly assume this theoretical attitude since there are important methodological and pragmatic reasons to act this way. The main reason behind this realistic assumption is precisely the need, mentioned earlier, to develop neuronal explanations with inductive power and projectability. John Bickle is clear about this inductive aspect when discussing the above quotation from Squire and relating it to evolutionary conserved traces:

> "This approach forges a connection between human neuropsychological data and experimental mammalian research. The 'particular structures and connections' namely the hippocampus proper, entorhinal cortex, perirhinal cortex, and perihippocampal gyrus, have homologs across the mammalian class. Since declarative (or explicit) memory is coextensive with hippocampal-requiring memory, the term is applicable to memory research on humans, other primates and rodents."
>
> (2003, p.78).

These considerations concerning the realist status of mental states and related explanatory relevance echo recent work on the philosophy of science and the topic of natural kinds. At the centre of this debate is precisely the way scientific practice organises and classifies reality in terms of (natural) kinds that, by means of their underlying structure, possess inductive relevance. A classical example is the modern chemical classification of substances in terms of their underlying micro physical properties.

In order to understand better the role of CN models and, in particular, the present discussion concerning realism of mental states, I will address some remarks made by Richard Boyd in his discussion of natural kinds. According to Boyd, the fundamental scientific practical act that establishes a certain term as denoting a natural kind is the process of *accommodation*, within a certain disciplinary matrix, between classificatory/taxonomical practices and 'real' causal structures. In the present discussion, within the specific disciplinary matrix which includes neuroscience, general biology, and evolutionary theory as background knowledge, we are able to state that we have find a natural (projectable) kind when we are capable to fit a certain mental state *f* with the appropriate (evolutionary preserved through species) neural-causal-mechanism M. Rephrasing this idea, Boyd claims that there are two different ways of defining a natural kind: a *programmatic definition* stating the functional role played by that kind within the disciplinary matrix, and an *explanatory definition* referring to the underlying causal properties that *justify* the functional role stated on its programmatic definition (Boyd, 1999, 70). In the context of CN models (assuming the general structure <B,*f*,M>), the *programmatic definition* of a natural kind *K* corresponds to the description of the role of *f* in explaining B whereas the *explanatory definition* consists in M's explanation of *f*. The natural kind *term* referring to the natural kind *K* is the mental-state term that fills *f* in a particular CN model. If we again take the spatial memory CN model as an illustration, spatial memory corresponds to the natural kind where its *programmatic definition* states that spatial memory causes "the optimization in spatial navigation of animals in their environment" and the corresponding *explanatory definition* (that justifies what is stated in the programmatic one) declares that spatial memory is explained by or corresponds to "The mechanism of LTP in Hippocampal Place Cells".

The important moral to be extracted from the discussion of mental states as natural kinds is that, in order to satisfy the projectability constraint, we have to consider mental states as real states that *explain* behaviours by virtue of their correlation to specific neuronal-causal mechanism. Anti-realist conceptions of mental states, in particular operationalist ones (functionalist and behaviourist), are ruled out from scientific practice since they are unable to answer the projectability demand.

2.1.2- *Hierarchical taxonomy of mental states*

The second feature concerning the nature of CN models consists of a qualitative clarification over the 'maximize projectability' norm. CN models vary among themselves in their 'grades of projectability' (or in their extensional scope within the domain **D**). Another way to state this characteristic is by noting that CN models come in different *grains of explanation* (Cf. Bechtel & Mundale, 1999). For instance, a model of contextual fear condition is *more fine-grained* and with a more restricted scope than a model of fear conditioning *simpliciter*. Hence, a consequence of this view is that models can overlap in the sense that a certain model can include a coarser-grained model (this happens in the former examples of contextual fear conditioning and fear conditioning). Here, again, CN models parallel Giere's approach to models in classical mechanics. For instance, a model of *Damped Pendulum* is simpler (less fine-grained) than a model of *Damped Driven Pendulum* and the latter includes the former.

In terms of CN models' general structure, this means that the grade of detail of the mechanist explanation M surely varies as the model is more or less coarse-grained. In any case, the mechanistic explanation of *f* will always be gappy or incomplete. Carl Craver (2007) calls these descriptions of incomplete mechanisms 'sketches', i.e. descriptions of mechanisms deploying black boxes or filler terms in them.[2] There are degrees of sketches, more or less incomplete. For my purpose here, what is important is that a sketch of a certain mechanism as deployed in a given model can be *filled in* by the adoption of a more fine-grained model that overlaps the former. So, the intended scope and grain of the models can vary depending on several kinds of qualification on mental states concepts, e.g. an overlapping/inclusive relation like the one between declarative memory and spatial memory, or as in composite relations like the one referred to above distinguishing (simple) fear conditioning from contextual fear conditioning.

The question of interest is that, when scientists intend finer-grained models, the mental state/process being modelled is not the same (e.g. fear conditioning is *not* the same mental state as contextual fear conditioning). The main lesson to be drawn is therefore that there is not a single 'level' of mental kinds (as is often suggested implicitly or explicitly) but a taxonomic *hierarchy* of mental states with different intended scope and projectability 'power'.

2.2- *The Cognitive Role and Status of CN Models*

One question that naturally arises is *how* these models are constructed or achieved. This section deals with some proposed (draft) ideas concerning this issue.

I ended the previous section with mental *concepts* when stating some of the most important features of CN models. In fact, we can understand these models (especially when regarding their cognitive role and status) as concepts, in particular

[2] Craver's approach concerns the process of neuroscientific discovery where, as more knowledge is achieved, more information is disposable to fill the blanks (or filler terms, or black boxes). My perspective, while not in conflict with Craver's is, nevertheless, distinct from his. In the CN models' framework the blanks in some mechanistic explanations (within a certain model) are not necessarily the result of scientific ignorance but, often, a practical methodological imperative taking into consideration the intended scope of the models.

mental/psychological concepts. For instance, a CN model for declarative memory can be regarded as establishing a *neuroscientific* explanation and theoretical/content fixation of the mental *concept* declarative memory. Ronald Giere suggests this direct link between his formal-semantic proposal and cognitive theories regarding the nature of concepts:

> "[A] model functions as predicate, as a model of a pendulum gives content to the predicate 'pendulum' in the open sentence '*x* is a pendulum'. So there is initially at least the possibility that some of what anthropologists, psychologists, and linguists have discovered about naturally occurred concepts might be carried over the study of the families of models[.]"
> (1999, p.100).

This rationale leads Giere to adopt a *Prototype Approach to Concepts* (e.g. Rosh and Mervins, 1975) to shed light on some cognitive features of his model-theoretical framework for classical mechanics. I follow Giere's suggestion concerning the relationship between models and the research on concepts but, contrarily to his endorsement of the Prototype approach, I will hold that the rival *Theory-Based Approach* (e.g. Murphy and Medin, 1985) is the most suitable for the present context of CN models.

In a very simplified way, the fundamental distinction between the Prototype conception and the Theory-based approach to concepts can be, very briefly and in a nutshell, stated as follows: the Prototype approach conceives concept formation and mastery as a matter of establishing sufficient sets of similarities between a certain object and a prototype. The prototype, as it were, defines a certain category; the object falls under that category if it shares a relevant set of similarities with the prototype. The Theory-based approach reacts against this view by claiming that there are virtually endless similar features between two objects or situations. A necessary condition to determine the right set of features considered relevant to establish an appropriate relation of similarity depends upon a consistent sets of beliefs or general theoretical knowledge that underlie our classificatory practices within a certain domain.

There is a deep relation between the theory-based approach to concepts and the topic of natural kinds (for an excellent discussion on this, see Griffiths, 1997, chapter 7). The central idea defining the Theory approach (concept formation depends on theoretical consistent knowledge) is basically the same as how to fix natural kind terms by relying on the process of accommodation within a certain disciplinary matrix. I also suggest that, in the context of CN models, mental states *f* should be considered as natural kinds whose causal-programatic definition is *justified* by an underlying neuronal-mechanistic explanations M. For these reasons (which I wont elaborate here) I tend to favour a theory-based approach regarding CN models.

The adoption of the theory view of concepts can be useful in discussion of how CN models are cognitively achieved by rephrasing this question as one concerning the issue of concept formation. Here I will adopt, tentatively, a recent cognitive approach to analogical reasoning that can shed light on this question. This proposed analogy is the so-called Multi-Constraint Theory of Analogy endorsed by Keith Holyoak and Paul Thagard (1997). One of the reasons for the adoption of this particular theory is that, although it is not explicitly stated, it seems quite consistent with the theory-based approach to concepts and with actual neuroscientific practice.

Very briefly, the multi-constraint theory of analogy states that analogical reasoning is achieved by the combination of three constraints in a situation where the source and the target analogues are put in relation: 1- similarity, 2- shared structure and 3- goal or purpose. This means that, according to this theory (and in parallel with the Theory-based account of concepts), "powerful analogies involve *not just superficial similarities*, but also *deeper structural relations*" (Thagard, 1996, p.81, emphasis added). Also, the purpose constraint puts the problem in the perspective of what it is intended to achieve and directs the similarity constraint to the relevant features to be reckoned as similar, given a particular context. This theory suggests that these three constraints act in parallel in order to achieve the best (analogical) solution to the problem at hand. I argue that this approach can somehow decompose the way concepts are formed within a theory-based account and, by extension, how CN models are cognitively and practically achieved.

Thagard and Holyoak (1997) suggest that when an analogy is successful some "induced explanatory schema" concerning the domain of application is achieved. I contend that, in the present context, besides fixating the content and extension of mental concepts, what is *inductively abstracted* from various neuroscientific analogies are CN Models adopting the general structure <B,*f*,M>. A glimpse of how this happens can, very crudely, be suggested as follows: given a certain (exemplar) animal experimental setting as the analogue source, the Purpose Constraint settles what behaviour B the experiment aims to study. The Similarity Constraint establishes the relevant behaviour similarities with other animal experiments (the analogy *target*) in order to understand B (as determined by the Purpose Constraint). The Shared-Structural Constraint establishes that a certain mental function *f* corresponding to a specific causal-neural mechanism M found in the source experiment must underlie (and therefore *explain*) the behaviour B in all the other (target) cases – using the evolutionary conservation principle as an inductive tool. The end result of this cognitive analogical process is an inductively-abstracted CN model <B,*f*,M>.

3- Ruthless Reductionism and Elimination: the *Memory Consolidation* Case

Recall now that Bickle's purpose is to show how current neuroscience provides reductions at the cellular-molecular level of certain mental states/processes (what he dubs Ruthless Reductionism). He gives the example of the supposed reductive Memory Consolidation Switch (MCS henceforth) – cAMP-PKA-CREB, Pathway link. Bickle also argues that this kind of accomplished reduction renders psychological explanation *otiose*: "When we have neurobiological causal explanations in place, psychological causal explanations are rendered *otiose*" (2003, p.114).

Using the CN model framework I will show that, actually, Bickle is making two very distinct claims. On the one hand he is claiming that (1)- MCS is *better* explained as *f* by a neuronal/componential (cellular-molecular) mechanistic explanation *Mc* than by a *competing* functional-cognitive/non-componential explanation. On the other hand, he is saying that (2)- MCS's explanatory role of a certain behaviour *B* becomes otiose once we get the reduction mentioned in (1). That is, once given the right explanation M of B we should drop *f* as unnecessary in the model; we would get the pair <B,M> instead of the triple <B,*f*,M> (M would *directly explain* B). While (1) deals with competing mechanistic explanations M of a certain mental state *f* and arguing for one of them (i.e. it argues for a particular *explanatory definition*), (2) states that the mental state *f* as such (as a *programmatic definition* – the role of explaining B) should be eliminated from the explanatory framework. It seems, thus, that (2) should be understood as a corollary of (1). Let's take a closer look at both claims.

3.1- *The vindication of claim (1)*

In relation to (1) Bickle is crystal-clear; there *was* a cognitive/functional explanation of MCS (the process of maturation from short-term memory to long-term memory) that became reductively suppressed by a more detailed and neuronal (cellular-molecular) successor. The previous explanation of MCS was based on psychological experiments and can be crudely summarised as follows: in order for a certain memory trace to consolidate from short-memory to long-term memory it has to be repeated *n* times during the relevant consolidation period without any retrograde interference. This would be the functional, information-processing and non-componential explanation of MCS. Bickle characterises this sort of explanation (within a structuralist model-theoretic framework of inter-theoretic reduction) as follows:

> "Models of a psychological theory of memory […] posit an entity/process, the consolidation switch. But they characterize this posit only in terms of the time course and amount of repetition needed to convert a type of memory item from short-term memory to long-term memory and the behavioral efficacy of different types of retrograde interference. In other words, psychology characterizes this entity/process in purely functional fashion"
>
> (Bickle, 2003, p.99).

He goes on to show how this purely functional explanation gets reduced by current cellular-molecular neuroscience, specially by the recent findings on the underlying cellular and molecular mechanisms of Long-Term Potentiation (LTP), in particular the mechanisms of molecular cascades that underlie the extension of Early Phase LTP (E-LTP) into Long Phase LTP (L-LTP):

> "The consolidation switch empirical base set in models of psychology got mapped to sequences and combinations of empirical base sets and fundamental relations in reduction-related models and intended empirical applications of molecular neuroscience, in particular to those involved in the transition of E-LTP into L-LTP and the maintenance of L-LTP. The empirical base sets of the latter include intracellular and neural transmission molecules: adenylyl cyclase, cAMP, PKA, CREB enhancers and repressors, DNA, RNA polymerases, ubiquitin hydrolase, CCAAT enhancer biding protein, glutamate, dendritic spine cytoskeleton components, AMPA receptors, NMDA receptors, and so on." (2003, 99)

So, basically, Bickle claims that the functional/information-processing explanation of MCS is replaced by a better neural (cellular-molecular) correlate. This process, Bickle claims, is a reduction, in particular a reduction of a certain previous functional/information-processing mechanistic explanation *Mf* of *f* (MCS in this case) for a new componential/neuronal one *Mc*.

The main and decisive reason to accept and endorse Bickle's suggestion for favouring Mc over Mf as an explanation for MCS (sustained, as it is, by neuroscientific practice) is simply that *within* the context of CN models what is required as mechanistic explanations M of mental states *f* are *neuroscientific* ones. If our goals are neuroscientific explanations, then it should not come as a surprise that, as a *matter of fact* and following *purely reasonable scientific practice*, a neuronal explanation should be favoured over a cognitive-psychological one. After all, we are searching for an *explanatory definition* of *f* within a disciplinary matrix where our background knowledge supporting inductive practices *is* a neuroscientific one.

In addition, it should be noted that Bickle's synoptic view of the place that the MCS – cAMP-PKA-CREB Pathway link holds in current neuroscientific practice is one excellent illustration of *what* a CN model is supposed to be and *how* it can be achieved. Recall that in the previous section it was stressed that, for explanatory reasons, a realist view concerning mental states should be adopted (which neuroscientists, in fact, do) and that realist assumption is satisfied by the direct reference to causal neuronal structures/mechanisms exhibited by terrestrial animals. It was also stressed that this realist assumption is what encourages neuroscientists to maximise the projectability of their explanations and classifications. In his report, Bickle shows us how current neuroscientific practice does this by seeking a causal-mechanistic explanation of MCS and by noticing, by experimental manipulation in different animal experiments, that the cAMP-PKA-CREB intraneural molecular cascade seems critical to *all* the instances of MCS (across behaviours and species). Bickle is particularly aware of the significance of this latter point in the projectability and inductive scope of this explanation:

> "There is a 'physical-chemical state' the cAMP-PKA-CREB molecular biological pathway, that uniquely realizes memory consolidation across biological classes, from insects to gastropods to mammals. […] These shared structures obtain despite vast differences in brain size, organization, site of principal effect (presynaptic or postsynaptic), behavioral repertoire, and even 'cognitive logic' of the distinct types of memory being consolidated (declarative versus nondeclarative)."
>
> (2003, p.148).

Bickle goes further and shows how neuroscientists use a phylogenetic conservation argument to cluster the several experimental results (in Mice, *Drosophila*, *Aplysia* and other species) under a single explanatory schema that justifies explanatory extrapolations and inductive conclusions. When all is said and done, what we get from John Bickle's example is a glimpse of the construction of a particular CN model (whose general form, recall, is <B,*f*,M>); a CN model of memory consolidation stating (tentatively) something like: <'Long time preservation of former acquired behaviors', MCS, 'cAMP-PKA-CREB Pathway>.

Within the CN model framework the adoption of the cellular-molecular neural mechanism instead of an information-processing one is, not only vindicated, but mandatory.

3.2- *The non-vindication of claim (2)*

Bickle thinks that, from claim (1), a more radical one follows directly: MCS's explanatory role of a behaviour B becomes otiose (i.e. dispensable) once a cellular-molecular mechanistic explanation of MCS is in place. More specifically, Bickle argues that there is a mental causation element implicit in that explanatory role that is swept away once this "new (cellular-molecular) kid is in town". The mental causation element is, in fact, clear if we tentatively state MCS's explanatory role as a *programmatic definition* within the CN model framework as something like: "MCS *causes* long-time preservation of former acquired behaviours". MCS *as a mental state* is postulated by its causal efficacy of a particular kind of behaviour (or so it seems).

Concerning mental causation (and MCS's mental causation in particular) Bickle states the following:

> "I contend that when we fix our gaze on aspects of scientific practice in this actual recent example, we see that psychological explanations *lose their initial status as causally-mechanistically explanatory* vis-à-vis *as accomplished* (and not just anticipated) cellular/molecular explanation. All attempts by philosophers to 'save' mental causation presuppose that psychological explanations remain (causally) explanatory."
>
> (2003, p.110, emphasis in the original).

That is, by achieving a low-level reduction of MCS, its overall functional/causal integrity would dissipate in the wider cellular-molecular explanation. In other words, (2) follows from (1). According to Bickle this is so *because* we can *now* provide a physical (cellular-molecular) causal-mechanistic explanation of the same behavioural phenomena MCS is supposed to cause/explain. After describing the cellular and molecular components of L-LTP maintenance (the cAMP-PKA-CREB Pathway), Bickle adds the following in relation to a complete cellular/molecular explanation of behaviour:

> "*Out of these components*, the fundamental relations of cellular/molecular neuroscience build membranes, plastic neurons, neuronal circuitries from sensory to motor effectors, neuromuscular junctions, and the specific connections with the muscles and skeletal systems (themselves built up by fundamental relations out of their molecular elements)."
>
> (2003, p.99, emphasis added).

In a more recent paper, Bickle's strategy is even clearer:

> "[T]he cellular or molecular events in specific neurons into which experimenters have intervened, in conjunction with the neuronal circuits in which the affected neurons are embedded, leading ultimately to the neuromuscular junctions bridging nervous and muscle tissue, *directly explain* the behavioral data."
>
> (2006, p.426, emphasis in the original).

So, Bickle thinks that a low-level explanation of MCS in terms of cellular/molecular mechanisms (the cAMP-PKA-CREB Pathway) *in conjunction with* an *embedment* of that mechanism *within a wider and detailed one* (at the same level), including a description of how sensory inputs and motor outputs are carried out at a very fine-grained level, provides an eleminativist reduction of the input-output functional relations characterising MCS since we got a better, more complete explanation: "The cellular/molecular neurobiological account explains many key causal processes that the psychological account is either completely blind to or leaves as input-output black boxes" (2003, p.113). It seems that (2) is a corollary of (1).

But there is something deeply wrong behind this reasoning. At the core of Bickle's contentions is the claim that we can provide a *direct explanation* of the target behaviours at the cellular-molecular level that could replace MCS's explanation as a mental psychological state. The problem with this proposal is that, in order for it to be achieved, we have to fill in the original mechanistic explanation M in the model of MCS (i.e. the cAMP-PKA-CREB Pathway) with *additional information* and complexity, as the quoted passages explicitly suggest. But this requirement leads Bickle to a dilemma! Basically, there are two incompatible ways of understanding Bickle's suggestion and both are unsatisfactory.

The first horn of the dilemma states that Bickle's solution precludes the very explanatory power of the model of MCS and the projectability of this mental state that vindicated claim (1) in the first place. This is so because the *detailed explanations*, including (*very detailed*) descriptions of the sensory and motor pathways involved in the particular behaviour to be explained, would be *very different* depending on the species and specific behaviours under study (as we saw in the previous section, Bickle *explicitly* recognises this wide variety). So, a *direct explanation of behaviour* as envisioned by Bickle would have to be a very detailed and concrete one, e.g. *Aplysia* gill-siphon sensitisation, or *Drosophila* avoidance behaviour, or mice orientation in the Morris-water maze, and so on. The undesirable consequence of this specification is that what was once clustered inductively as one explanatory schema is now fragmented. These *direct* explanations would be topical, token-specific, behaviour-specific and species-specific. Bickle's strategy of filling the very gappy cAMP-PKA-CREB Pathway mechanistic sketch with additional information has the consequence of creating *new* very specific and fine-grained CN models *overlapping* the MCS one. That being so, these new models *would not* be CN models of MCS any more but much more specific fine-grained CN models of very diverse mental states, *specifying* at a very detailed level the memory consolidation process in particular cases like *Aplysia* gill-siphon sensitisation, *Drosophila* avoidance behaviour, or mice orientation in the Morris-water maze. So, Bickle's proposed *direct explanation* of the behavioural data has the chief unpleasant consequence of a strong limitation in scope of those direct explanations. But this is a consequence that Bickle himself explicitly avoids in supporting claim (1). So, in fact, not only does claim (2) not follow from claim (1) but it is incompatible with it. So, the first horn of the dilemma states that an explanation formulated in terms of *physical causation* instead of *mental causation* would have to pay the price of explanatory irrelevance.

The only possible way off the first horn of the dilemma leads us to the second one. It follows from maintaining that MCS is, in fact, the function to be explained and instead of its multiple particular instances. In other words, what is sought is a CN model of MCS and not of its possible overlapping models. But, again, (for the same reasons stated above) we *cannot* provide a detailed and fine-grained description of the whole process in the model of MCS if we want to keep it *empirically correct*, i.e. as *intending* to cover *all* the elements of the general domain – terrestrial animals – satisfying the open sentence “*x*Displays MCS” (respecting the maximise projectability constraint). There is too much variation in how the different species realise the cAMP-PKA-CREB molecular cascade (pre-synaptic, post-synaptic, etc.) as well as in all the implementational tissues, from the sensory receptors to the motor pathways involved in *all possible* behaviours in *all possible* species. In *this* sense, MCS is *multiple realisable*. A purely neural-causal explanation of MCS is therefore not possible after all. In this case we would only guarantee explanatory relevance at the cost of having to assume mental causation and disregard a possible physical (causal) alternative. This is the second horn of the dilemma.

Recognition of this dilemma leads again to the initial problem that motivated Bickle’s claim (2) in the first place. It seems now that an answer to the mental causation challenge is not possible in the way envisioned by Bickle. Does this consequence means that the (naturalistic rephrased) classical/metaphysical argument in favour of the ‘reality’ of mental causation based on the assumption of the multiple realisability of mental states is fundamentally correct? I will address these questions in the next section

4- ‘Functional Explanation’ vs ‘Mental Causation’ and Reduction without Elimination

4.1- *Functional Explanation vs Mental Causation*

A first step overcoming the above stated hindrance is to take into consideration, in the present discussion, the very nature of *mechanistic explanations*, as those required in CN models are conventionally represented by the letter M. Generally, a mechanistic explanation of a phenomenon *R* is a description of the (relevant aspects of the) mechanism that produces that phenomenon. More specifically (and adopting Carl Craver’s terminology in several places) this explanation depicts a (finite) set of *components* (Φs) and *activities* (Ωs) organised in such a way as to produce the *role* or *effect* (*Ψ*). In the present context, it should be emphasised that, as a matter of fact concerning the nature of mechanistic explanations, the role or effect *produced* by the cAMP-PKA-CREB Pathway molecular cascade is *not* “long time preservation of former acquired behaviors” (the target behaviour in the model that MCS is supposed to cause) but merely *dendritic growth*. The link to memory consolidation is a further step.

What allows us then to relate the cAMP-PKA-CREB Pathway mechanism to the effects on the consolidation of behaviours? Basically, this relation is established by framing this mechanism within an explanatory CN model of MCS. More precisely, the link is established when, *within* this model/explanatory framework, scientists manipulate at the level of the mechanism in certain experimental settings and achieve certain behavioural results concerning the long-time preservation of former acquired behaviours. This is precisely what Bickle reports in some examples of the 'intrevene molecularly, track behaviourally' strategy, such as the experiments dealing with CREB-knock-out and CREB-mutant mice where (in both cases) the gene-affected animals were unable to display long-time preservation of some former acquired behaviours measured in several experimental protocols (fear-conditioning, social recognition, spatial memory, etc.). If to this we add Carl Craver's adoption of the neuroscientific practice of Woodward's conception of causation as manipulation, we are safe to assume that cAMP-PKA-CREB Pathway *causes* 'long-time preservation of former acquired behaviors', where *causes* is as physical as you can want and get. Considerations for or against this proposal aside, it should be stressed anyway that if we adopt the naturalistic/pragmatic stance (Bickle's New-Wave Metascience) this *is* the kind of causation holding a great deal of contemporary mainstream neuroscientific explanations (cf. Craver, 2007, especially chapter 3). Scientific practice itself settles the question, disregarding purely metaphysical considerations.

Now, it is easy to show how the multiple realisation problem faced by Bickle's approach emerged and how it can be avoided. Contrary to Craver, Bickle sustained his argument on the assumption of a notion of causality oddly close to a classical metaphysical conception of Physical Causation according to which an Event1 physically causes an Event2 only if some *physical property* is transmitted from Event1 to Event2. This is precisely the assumption challenged by the manipulationist conception of physical causation. When Bickle proposes the *direct causal link* between a certain neural mechanism M and a behaviour B he is thinking in terms of a chain of transmitted physical properties, and it is this view that leads him to claim the necessity to add information concerning the embedment of that mechanism within a wider and detailed mechanism and specifying how specific sensory inputs relate to specific motor outputs. As we have seen concerning the specific case of MCS, *this conception* is faced with a dilemma: either we ask for a physical-causal explanation and face a lack of explanatory relevance, or, if we insist on explanatory relevance we face multiple realisation. The only way to avoid the dilemma is to embrace mental causation instead of physical. The adoption of the manipulationist account overcomes this problem since no transmitted properties are required to establish soundly a physical causal relation between two events. *Very* roughly, this conception generally states that if one intervenes in a certain way and in certain conditions in (some of) the properties of Event1 and that intervention has an effect on (some of) the properties of Event2, then we are able to claim, in a certain context, that Event1 *physically causes* Event2 (this is more a caricature of the manipulationist conception than a rigorous and technical description, but I lack the space for the more complete analysis that this approach deserves).

So, *within this* CN model, the cAMP-PKA-CREB Pathway neuronal mechanism and MCS, as a mental state, are *causally isomorphic*, i.e. they are extensionally equivalent in the schema "'X' *causes* 'Long time preservation of former acquired behaviours'". This conclusion answers the above-stated anxiety concerning the nature of MCS's causal status. There is not an inevitable autonomous and intrinsic *mental* causation linking MCS as a mental state *f* and the target behaviours B. This causal relation is more an heuristic statement than substantial/metaphysical one since within the CN models framework it is assumed that neuroscience's aim is to find the neural/physical correlates of that relation. Rephrased in Boyd's terms, MCS's explanation of consolidation of behaviours (in terms of a causal relation) is a *programmatic definition* of MCS as a natural kind, which holds *because* there is an *explanatory definition* relating MCS to the cAMP-PKA-CREB Pathway. In face of these considerations, instead of the metaphysically-charged expression Mental Causation, one should use the phrase Functional Explanation – stressing its *programmatic* status – when regarding the role of MCS (as a mental state) in the context of MCS's CN model.

4.2- *The Methodological Preservation of Functional Explanations*

The above considerations lead to an important conclusion concerning the *non-eliminable* status of mental states generally. MCS maintains its functional integrity in part because, as we have seen, it is its functional profile that frames the mechanism explanatory significance in the first place. It is MCS's programmatic role that settles the cAMP-PKA-CREB Pathway's explanatory significance. Again, Boyd emphasises that in relation to natural kinds generally "their 'explanatory definitions' *explain why* they satisfy the 'programmatic definitions' (1999, p.71, emphasis added). Among other things, MCS (as a mental state) establishes the Purpose Constraint in the analogical reasoning underlying the CN model formation (therefore, helping to sort out the *relevant* similarity set of features that cluster the different empirical evidence together under the same explanatory schema – namely the long time preservation of former acquired behaviours caused by MCS).

Bickle can, nevertheless, agree with these last remarks. His response would be that what happens in real practice is that once this programmatic role is accomplished by a suitable physical neuronal explanation, its heuristic status is rendered *otiose* and abandoned henceforth. This is Bickle's contention, after all. But this conception is committed to a strange view of scientific discovery process. In particular, a view that seems to prohibit or rule out any kind of scientific objection, doubt or alternative proposal regarding a certain neuroscientific reduction or explanation of a certain mental function/state. This is, no doubt, an undesired consequence of Bickle's eliminativist stance. In fact if we accept as obvious that *there is no final explanation*, some methodological autonomy is necessary if we are able to revise proposed mechanistic explanation once new data are available. I will provide an actual example from current neuroscience that illustrates this. The example is not related to that parcel of neuroscience I have been calling cognitive neuroscience (i.e. explanation having as its target explanation of creatures' behaviours). The positive side of this is that it illustrates that functional role heuristics is not the sole privilege of CN explanations. The example I have in mind concerns the Retrograde Information posited within the wider explanation of Long-Term Potentiation.

Very briefly, the idea of Retrograde Messenger can be summarised as follows. Some neuroscientists suggest that LTP, being *induced* and *expressed* post-synaptically, is also expressed pre-synaptically, namely by producing more vesicles containing neurotransmitters (notably glutamate). But, in order for the pre-synaptic cell to react this way, some form of information from the post-synaptic cell must travel backwards against the usual unidirectional synaptic signalling (from the pre-synaptic cell to the post-synaptic one). Accordingly, neuroscientists posited that some form of retrograde messenger/signalling must occur to send information back from the post-synaptic to the pre-synaptic cell. The retrograde messenger corresponds to a *functional defined* role which *explains* the pre-synaptical expression of LTP. The next step in this explanation consists of suggesting and finding a chemical realiser of this functional role and corresponding physical mechanism concerning how this putative substance is produced and how it acts in order to produce the target phenomenon.

In the early nineties some authors proposed that nitric oxide (NO) (discovered, at the time, as an endogenously-produced substance) could *be* the retrograde messenger. Some (incomplete and hypothetical) mechanisms were proposed to explain how NO was synthesised in the post-synaptic cell, how it could reach the pre-synaptic cell and how it could contribute (indirectly) to the expression of LTP there. At this point, it could be helpful to suggest a model to frame the situation described. The only difference regarding CN models consists of replacing the *B* in the general structure by, say, *P*, standing for the phenomenon to be explained. In this case, the triple <P,*f*, M>, could state something like: <Presynaptic LTP expression, Retrograde Messenger, NO (and specified mechanism)>. As it happens (and happens constantly in science) the claim that NO is the retrograde messenger has been disputed and other substances have been suggested as realising that function (e.g. endocannabinoids). The important point to be emphasised in the present context is that the suggestion that NO could assume the role of the retrograde messenger *did not* render the 'pure' functional (*heuristic*) characterisation of retrograde messenger *otiose*. The very suggestion that other substances can realise that same function would not be possible if some autonomy of that functional (and therefore heuristic) role did not exist. *Mutatis mutandis*, and by the same token, the heuristic role of mental states functionally defined is not precluded in CN models by a proposed mechanistic/neuronal explanation.

5-Conclusion

I used the framework of CN models to try to make clear that cognitive-neuroscientific explanations *have to* obey two norms in order to fulfil some basic explanatory competences (e.g. inductive power). The first is that, assuming the proper disciplinary matrix in which CN models are constructed, mental states *f* explanations of behaviours B (the *programmatic definitions*) are justified only on the *background assumption* that they are, *in turn*, explained by/reduced to neuroscientific mechanistic explanations M (the *explanatory definitions*). Secondly, mental states are not eliminable once a supposedly good neuroscientific explanation is in place; they preserve a fundamental heuristic/programmatic role established by their functional profile that sets what is to be explained by neural-causal mechanisms. Of course, mental states' functional role *can* be revised in the light of empirical data (like the example of the splitting of Memory into more specific kinds) but the new kinds are *still* mental states defined functionally and *still* preserve a programmatic role. In short, within the disciplinary matrix of current cognitive neuroscience, functionally-defined mental states *f* without a correlated neural-causal explanation M are *empty* and neural mechanistic explanations without functionally-defined mental states are *blind*.

References

Bechtel, W. & McCauley, R. N. (1999). Heuristic Identity Theory (or Back to the Future): The Mind-Body Problem Against the Background of Research Strategies in Cognitive Neuroscience. Proceedings of the 21st Annual Meeting of the Cognitive Science Society, pp. 67-72.

Bechtel, W. & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. Philosophy of Science, 66, 175-207

Bickle, J._(2003). Philosophy and Neuroscience. (Kluwer Academic Publishers)

Bickle, J. (2006). Reducing Mind to Molecular Pathways: Explicating the Reductionism Implicit In Current Cellular and Molecular Neuroscience. Synthese, 151, 411-434

Boyd, R. (1999). Kinds, Complexity and Multiple Realization. Philosophical Studies, 95, 67-98

Craver, C. (2007). Explaining the Brain. (Oxford University Press)

Giere, R. (1988). Explaining Science. (University of Chicago Press)

Giere, R. (1999). Science Without Laws. (University of Chicago Press)

Griffiths, P. (1997). What Emotions Really Are. (University of Chicago Press)

Holyoak, K. J. & Thagard, P. (1997). The Analogical Mind. American Psychologist, 52, 35-44

Murphy, G. & Medin, D. (1985). The Role of Theories in Conceptual Coherence. Psychological Review, 92

Rosch, E. & Mervis, C. (1975). Family Resemblance: Studies in the Internal Structure of Categories. Cognitive Psychology, 7, 573-605

Shouten, M. & de Jong, H.L. (Eds.) (2007). The Matter of the Mind. (Blackwell)

Squire, L. (1992). Memory and the Hippocampus: A Synthesis From Findings with Rats, Monkeys, and Humans. Psychological Review, 99(2), 195-231

Thagard, P. (1996). Mind: Introduction to Cognitive Science. (MIT Press).

Woodward, J. (2003). Making Things Happen. (Oxford University Press)

Wright, C. (2007). Is Psychological Explanation Becoming Extinct? (In M.Shouten & H.L.de Jong (Eds.) The Matter of the Mind (pp.249-274). Blackwell)