Logique & Analyse **135-136** (1991), 297-318

CAUSAL EFFICACY, CONTENT AND LEVELS OF EXPLANATION

Josefa TORIBIO

1. *Paradoxes and Dilemmas*

Let's consider the following paradox (Fodor [1989], Jackson and Petit [1988] [1992], Drestke [1988], Block [1991], Lepore and Loewer [1987], Lewis [1986], Segal and Sober [1991]):

i) The intentional content of a thought (or any other intentional state) is causally relevant to its behavioural (and other) effects.
ii) Intentional content is nothing but the meaning of internal representations. But,
iii) Internal processors are only sensitive to the syntactic structures of internal representations, not their meanings.

Therefore it seems that if we want to defend the idea — absolutely plausible from an intuitive point of view — that mental / intentional states are causally responsible for behavioural outputs and we want to do it on the physicalist basis of any scientific methodology, we will have to give up the conviction that such intentional states *qua* intentional, *i.e.* as having a particular meaning, are the ones causally responsible for our behaviour.

The path that takes us to mental epiphenomenalism is clear: 1) the causal powers of any event are completely determined by its physical properties; 2) although intentional properties supervene on physical properties, they can't be identified with them; 3) intentional properties, as intentional, are not causally responsible for behaviour, because they don't take part in the causal powers of the states to which they belong, *i.e.*, intentional properties are *epiphenomenal*.

Let's consider now a different yet parallel position to the one just described. There is an important debate in cognitive science about whether the class of mechanisms to which we belong and which the computational modelling project of cognitive processes refers to is best represented by classical or connectionist approaches (McClelland, Rumelhart *et. al* [1986], Smolensky [1987] [1988], Fodor and Pylyshyn [1988], Pinker and Prince [1988], Clark [1989], Ramsey, Stich and Rumelhart [1991], Clark and

Karmiloff-Smith [*forthcoming*]).

In classical, serial processing models, information is encoded in terms of rules that have a linguistic character. In connectionist or parallel distributed processing (PDP) models, the causal relationships among the units that constitute the system determine how the information is processed by the network, although these units don't have a *direct* semantic interpretation. But, for both, classical and connectionist models, all the computations can be explained without any reference to the *content* of the processed information, *i.e.*, in both cases the properties that seem to be responsible for the system's behaviour are ultimately physical properties nor intentional ones. This situation mirrors thus, within cognitive science, the philosophical discussion concerning the causal efficacy of semantic properties.

Now, if in this debate we opt for the classical paradigm, there is a way of finding a solution to the computational version of the epiphenomenalism paradox. This solution is based mainly on the notion of supervenience or, more precisely, on the notion of mereological supervenience (Kim [1984] [1988])([1]). The idea of intentional properties supervening on physical properties makes sense within the classical context because there exists an easily isolable supervenience base comprising the syntactic items in the so-called language of thought.

But, what happens if we opt for the connectionist paradigm?. The situation here doesn't seem to favour the use of the same supervenience strategy. For it has been argued (Ramsey, Stich and Garon [1991]) that beliefs, desires, and other mental states are nor, in the connectionist paradigm, individuable as weight or activation states of the system. This is because information is encoded by the network in distributed and superpositional representations, i.e., there are no straightforwardly isolable vehicles at the physical level that can be identified as the articulated supervenience base on which the semantic

([1]) The notion of supervenience was originally introduced by G. E. Moore in his characterization of the relations between evaluative and descriptive properties. That evaluative properties supervene on descriptive properties means that there is no difference in the former without a difference in the latter. In this general sense, the notion of supervenience is a variation of the general thesis that states that physical facts determine all the facts. In section 4 I'll offer a more detailed treatment deriving from the influential work of Jaegwon Kim.

properties supervene(2).

If this is true, then the connectionist not only loses the battle against epiphenomenalism but more drastically, seems to offer a standing invitation to eliminativism, since talk of beliefs and desires, etc. now seems to be floating free of any acceptable scientific underpinning, i.e., she has lost the necessary theoretical apparatus for supporting the intuitive idea that propositional attitudes — beliefs, desires and any mental states with semantic content — are physically realized. This second line of argumentation doesn't take us to a paradox but to a dilemma: either we accept eliminativism, if connectionist hypothesis are correct or we defend the causal efficacy of mental states with semantic content by showing that, after all, connectionist networks are not plausible cognitive models (Davies [1991]).

From my point of view, however, both lines of argumentation need to be revised. The aim of this paper is to find an account of the causal efficacy of content that avoids the aforementioned epiphenomenalist objections and that doesn't require the discovery of inner symbols in the computational modelling of such contentful mental states. In short, the aim is to find a meeting point where a philosophical story about content and cause and the connectionist computational model can be brought together (*Cfr.* Clark [1989]).

2. *Relevance versus Efficacy*

How can we reconcile the fact that semantic level properties seem to be explanatorily *relevant* but not causally *efficacious* — inasmuch as the causal power of the states that have them lies in the micro-physical states that realise them?.

This problem involves a discussion of some of the most representative positions in cognitive science, basically those lumped together under the labels of the *Representational Theory* (Fodor [1975] [1987]) and the *Syntactic Theory* (Stich [1983], Churchland [1986] 1989]). The Representational Theory can roughly be characterised as the attempt to

(2) A distributed representation is one constituted by a set of microfeatures, which are subsymbolic, *i.e.*, which are not semantically interpretable themselves. When the same unit or units that contribute to encode information involved in one representation also contribute to encode information involved in many other representations, we say that the representations are superpositional.

conceptualise mental states in terms of their relationships with some kind of representational entities — sentences written in some mental code or language of thought — that are then conceived as semantically interpretable and causally efficacious in virtue of their semantic content. Classic models in AI fit this kind of story nicely.

The core of the Syntactic Theory is, on the contrary, the idea that cognitive states can be systematically projected onto abstract syntactic objects, in such a way that the causal chains between stimuli and behavioural events can be described exclusively in terms of the syntactic properties and relations of these objects, without any need to appeal to their semantic content. It shouldn't be a surprise then that the connectionist models appear in this case as providing the empirical support for defending the central thesis of the Syntactic Theory.

An exhaustive review of this philosophical / computational landscape, even one restricted to the specific topic of mental causation, is beyond the scope of this paper. My aim is much more modest. I intend to argue for the possibility of intentional laws with the help of conceptual tools which belong to the computational realm. After all — and although this is not a thesis free of controversy — the question about whether a property P — semantic or not — is causally responsible for certain behaviour comes down to the question of whether or not there are causal laws involving P([3]). Therefore, if it can be shown that there are causal laws covering semantic properties and if it can be done without appealing to any mental code or language of thought à la Fodor, the twofold aim will have been reached.

According to the preceding thesis, P is a causally responsible property if it is a property in virtue of which the individuals that instantiate it can be subsumed by causal laws. In order to be able to correctly hold that an event c has caused an event e there must be some properties F and G such that c instantiates F and e instantiates G and "'F instantiations are sufficient for G instantiations' is a causal law" (Fodor [1989], p. 64).

The ontological commitment underlying this thesis is the following. Only individual events can be causes. But, at the same time, the necessary character of the regularities expressed by a causal law depends on such regularities being established not between particular events, but between

([3]) Davidson's influence in the way of addressing this question should be clear. His anomalous monism, *i.e.*, the idea that mental events come into causal relationships only under their physical description because this is the only description in which they are subsumed by laws, involves the aforementioned thesis.

types of events. Now, since particular events can be referred to by many different expressions, some of which don't mirror any of the properties that turn them into causes or effects of other events, the criteria for grouping together particular events into events of the same type — events of the type that can appear in a causal law — must only focus on those properties that can be shown to be causally efficacious.

For instance, although it might be true that the dinner on Sunday caused my stomach ache, the causal law on which such truth depends doesn't state any relation between events of the type *to have dinner on Sunday* and *stomach ache*, but rather between events of a particular physical type, such as a certain composition of the meat and an alteration in the digestive juices.

In short, what makes a causal law a proper law, and not a mere regularity with enough statistical support, is the existence of a micro-physical mechanism or structure that is shared by the different macro-types of events that are subsumed by that law. In other words, a macro-type of events can be considered causally efficacious only if it supervenes on micro-physical events — perhaps different on different occasions — in such a way that the causal powers of the former are *explained* by the causal powers of the latter.

When the matter is to individuate causal laws, it seems as if, strictly speaking, there were only micro-physical laws. The rest — all the laws of the so-called special sciences, such as Social Sciences, or the intentional *laws* of Psychology, based on the semantic content of mental states, are only approximate formulae that inevitably include *ceteris paribus* clauses.

The distinction suggested by Jackson and Pettit (Jackson and Pettit [1992]) between causal *efficacy* and causal *relevance* has its root in the same kind of considerations. The only difference is that what they call *causal* relevance is, from my point of view, an unlucky term to designate what I've named here *explanatory* relevance.

In effect, for Jackson and Pettit, both the causal stories at an intentional level and the causal laws with which the special sciences work are only *causally* relevant (or, in my terminology, *explanatorily* relevant), while the proper *causal efficacy* lies exclusively in the more basic micro-physical level. Their line of argument is similar to the one stated above. Only individual events described from a micro-physical point of view can be causes. Now, causal stories at a higher level don't unfold in terms of singular events, but generalise over conjunctions and disjunctions of lower level events. Therefore, the causal laws that belong to these higher levels can only be true in virtue of the causal efficacy of the micro-physical events which the terms of the macro-physical or intentional vocabulary that appears

in those laws refer to in the end.

Although persuasive, this line of thought places its profits into the epiphenomenal account that we wanted to settle. It can be granted to the realist scientist that nothing exists but the entities described by physics, but that doesn't mean that we should devalue the causal status of those laws that don't involve such entities. That would entail a misunderstanding of the very character of the scientific methods and explanations. In the next section, I try to resist this kind of argumentation through the defence of an anti-reductionist thesis.

3. *Strict Laws and ceteris paribus Laws*

Many philosophers now believe that the reductionist program has proved to be misguided (a good vindication of this anti-reductionist line can be found in *The Scientific Image* by Van Fraasen). Why should the semantic properties of our mental states appear at a physical level?. Nobody expects to find them there. Thousands of objects of unquestionable existence — tables, mountains, hurricanes ... — don't exist in the micro-physical vocabulary and yet they aren't devoid of their causal powers for that reason. The laws that let us predict the formation of a hurricane in the Pacific or that let us explains the consequences of such a formation, once they are reduced to a purely micro-physical vocabulary don't mention any entity that can still be called a hurricane. And yet, sciences such as geology, meteorology and other equally respectable base their explanations in the existence of causal laws that connect phenomena described in the vocabulary of those sciences.

Especially relevant is the case of biology or chemistry. It is not the case that biologists or chemists hold that there is something beyond the physical processes underlying biological or chemical processes. It is just that explanations using terms as *acid* or *alkaline*, or *variation* and *selection* are good explanations in virtue of the existence of laws that relate properties whose causal efficacy only makes sense in the vocabulary of chemistry of biology.

The thesis that beliefs, desires and other mental states work as internal causes of behaviour and do it so in virtue of their semantic content can be justified by using the same kind of argument that has been used for the special sciences case. The desire for a cold beer plus the belief that the beer is in the fridge caused Andy to go to the kitchen and open the fridge. Or,

in general,

(x)(p)(q) [(x desires p) → (x believes (q → p)] → *ceteris paribus* x brings about that q

This has been, for instance, the line of argumentation defended by Fodor in two important papers: "Making Mind Matter More" (Fodor [1989]) and "You Can Fool Some of the People All the Time, Everything Else Being Equal: Hedged Laws and Psychological Explanations" (Fodor [1991]).

In the first, Fodor shows how the legitimacy of causal laws formulated in terms that belong to a higher level than the micro-physical one is a problem that arises not only in psychology but in any of the *non-basic* or special sciences. This problem, however, doesn't seem insoluble if we notice that the only difference between basic and non-basic causal laws is that, in the latter, there has to be a *mechanism* in virtue of which the satisfaction of the antecedent guarantees the satisfaction of the consequent or, in other words, there has to be a mechanism that implements such laws. And the main point is: although the mechanisms that implement intentional laws — laws covering the content of mental states — have a physical character, the laws in themselves are essentially intentional and justify the ascription of causal efficacy to such states insofar they are semantically interpretable.

Now, the existence of physical mechanisms of implementation in the case of non-basis laws, turns them into non-strict laws. These laws always involve *ceteris paribus* clauses and, therefore, the question that immediately arises is how it is possible to guarantee the ascription of causal efficacy to the semantic properties of the states subsumed by intentional laws when they unavoidably involve those clauses.

The problem gets even worse if, on the basis of the multiple realizability of intentional states, we argue — as Schiffer does (Schiffer [1991]) — that the very notion of *ceteris paribus* law doesn't make any sense in psychology, among other reasons, because it is always possible to find significant exceptions to those laws, so significant as to make it impossible to determine their truth-conditions([4]).

([4]) Shiffer's position is, briefly that i) the defence of intentional laws is absolutely implausible once the multiple physical realizability of mental states is taken into account and that ii) nevertheless, the explanatory power and predictive character of intentional theories don't depend at all on the existence of such laws with or without *ceteris paribus* clauses. The second of these two thesis is also held by A. Clark, although his line is more

Fodor's answer to this problem, developed in the second paper mentioned above, is complex and requires a certain amount of special terminology, but roughly it comes down to the following argument:

a) It is assumed that any type of intentional state (*A*) in virtue of which an organism satisfies the antecedent of a *ceteris paribus* causal law is a functional state whose physical realization can be different in different organisms or in the same organism at different times.

b) The fact that intentional laws incorporate *ceteris paribus* clauses shows that the causal efficacy of a type of intentional state (*A*) with regard to a certain behaviour (*B*) requires the joint existence of some of the possible physical realizations of that state *and* the additional conditions (*C*) that are clustered together under those clauses. In that way, if R1 is a realization of the event type *A*, such additional conditions would be represented by an arbitrary type of event *C* if:

i) *A*(R1) and *C* are (strictly) sufficient to bring *B* about.
ii) It is not the case that only *A*(R1) is sufficient to bring *B* about.
iii) It is not the case that only *C* is sufficient to bring *B* about.

c) According to this formulation, if R1 is a physical realization of the state A, there are three different ways in which the presence of R1 could involve, however, exceptions to the law "*ceteris paribus* A → B" :

i) First, when the additional conditions (*C*) that are necessary for the production of *B*, although determinable, don't take place *jointly* with the realization of R1. We have then what Fodor calls a *mere* exception.

ii) Secondly, when there are no such additional conditions, i.e., when there is no event type *C* such that occurring jointly with the realization R1 of A constitutes a sufficient condition so as to bring *B* about. In this case, all tokens of *A* realized by R1 are *absolute* exceptions to the law.

iii) Finally, when the realization R1 of *A* is an absolute exception not only in regard to the law "*ceteris paribus* A → B" , but in regard to all laws in which A is the antecedent ("*ceteris paribus* A → D" , "*ceteris paribus* A → F" , etc).

behaviouristic in character. Cfr. Clark [1993].

According to Fodor, what distinguishes strict laws from *ceteris paribus* laws is that the latter may have exceptions, even absolute exceptions in the sense explained in ii). What distinguishes *ceteris paribus* laws from empty propositions is that in the former, but not in the latter, the physical realizations that correspond to the intentional states subsumed by the antecedent are not absolute exceptions to all laws in which the same antecedent appears, i.e. they are not absolute exceptions in the sense explained in iii).

The final justification of why such exceptions are not possible, i.e. the final justification of the nomological character of the *ceteris paribus* laws, despite the existence of the other types of exceptions, lies in the fact that, in the case of intentional laws, the notion of physical realization is defined *functionally*. Once this functional character is assumed, any physical state that was an absolute exception to the set of laws in which an intentional state A appears as antecedent, couldn't simply be individualized as a realization of that intentional state, because there wouldn't be any external criterion that let us determine what state we are talking about. The problem, then, is not so much a problem related to the possibility of *ceteris paribus* laws but rather a problem related to the correct functional individuation of the states that are covered by those laws.

Let's see how this strategy works in a concrete case. Let's imagine there are only three laws in which the state "wanting to lose weight" appears as the antecedent. These laws are:

- *ceteris paribus* people who want to lose weight put themselves on a diet.
- *ceteris paribus* people who want to lose weight take some exercise.
- *ceteris paribus* people who want to lose weight utter sentences of the kind "I'd love to lose a few pounds" .

Furthermore, let's suppose that one of the physical realizations proposed for that state is to be in a neuronal configuration S and, finally, let's imagine that people who happen to be in that kind of configuration are absolute exceptions to the three aforementioned laws, i.e., they are absolute exceptions to *all* the laws that involve the mental state "wanting to lose weight".

As we've already seen, that means not only that, now and then, people who want to lose weight don't put themselves on a diet, or don't make any exercise, or don't utter the sentence in question. It rather means that there are not additional conditions *C1* such that if someone was in the neuronal

configuration S and those conditions were met, then she would put herself on a diet, *and* there are no additional conditions *C2* such that if someone was in the neuronal configuration S and those conditions were met, then she would take some exercise, *and* there are no additional conditions *C3* such that if someone was in the neural configuration S and those conditions were met, then she would utter "I'd love to lose a few pounds".

In this case, i.e., when there are no additional conditions that *complete* the state of being in a neuronal configuration S with respect to *any* of the laws in which "wanting to lose weight" appears in the antecedent, we have good reasons, not for questioning the nomological character of our generalizations, but to reject the idea that "to be in S" is the physical realization of "wanting to lose weight". This is so because, after all, what defines the mental state "wanting to lose weight" is its function, and its function is to cause at least some of the behaviours mentioned in our example. In this way the only exceptions that might get the process of validating non-strict laws into trouble are ruled out.

From my point of view, however, Fodor's defence of the legitimacy of *ceteris paribus* laws is, in some sense, circular([5]). The reason is the following. If the functional role of a representation or mental state is — at least according to a broad functionalist point of view — its causal role, a defence of *ceteris paribus* laws carried out as a way of guaranteeing the ascription of causal efficacy to the semantic properties of those states that rests, in the end, in their correct functional individuation (*i.e.* in their correct *causal* individuation) doesn't seem to be the best of the possible defences. An equivalent argument, and equally circular, would be one that vindicated the validity of the law "*ceteris paribus* if it cuts glass, then it's a diamond" by appealing to the fact that the functional individuation of what it is to be a diamond includes the property of being able to cut glass.

There is however an argument in favour of the legitimacy of *ceteris paribus* laws that doesn't involved those problems. It is based on a definition of causation in terms of counterfactuals that has its origin in D. Lewis (Lewis [1986]) and on the vindication of a certain notion of supervenience that not only lets us account for the multiple realizability of the micro-physical structures underlying the same types of intentional states, but also

([5]) This "sense", which I will immediately develop, is different from Schiffer's criticisms about circularity. Schiffer's criticisms are, I think, satisfactorily answered by Fodor at the end of this paper. Cfr. Fodor [1991], pp. 31-33.

does so by establishing a nomological relationship between the states so characterized. I turn now to this alternative approach.

4. *Counterfactuals and Supervenience*

In its most basic formulation, a definition of causality in terms of counterfactuals can be given in the following way:

> "If *c* and *e* are two actual events such that *e* would not have occurred without *c*, then *c* is a cause for *e*" (Lewis [1986], p. 167).

Of course, this definition is in need of important provisos. Let's suppose, for instance that a bright red piece of coal causes my cigarette to be lit. We might say that had the piece of coal not been bright red, the cigarette wouldn't have lit. But, although that counterfactual is true, it doesn't establish any kind of causal relationship between the properties of being bright red and being lit.

If we want to guarantee that the macro-properties used to describe events counterfactually connected in a putative causal explanation are really efficacious with respect to the described behavioural outputs, we need to establish some kind of nomological connection between those macro-properties and the mechanisms that ensure their causal efficacy. That kind of connection is provided by the notion of supervenience or, more precisely, by the notion of mereological supervenience. This notion is expressed in its most general form by Kim in the following terms:

> "*...the supervenience of a family A of properties on another family B* can be explained as follows: necessarily, for any property F in A, if any object x has F, then there exists a property G in B such that x has G, and necessarily anything having G has F. When properties F and G are related as specified in the definition, we may say that F is supervenient on G, and that G is a supervenience base of F" (Kim [1984], p. 262).

Kim uses the notion of supervenience as his basic tool to argue against the epiphenomenal treatment of mental properties. Mental causation is nothing but a case of supervenient causation and, as such, the semantic properties of the mental states that come into supervenient causal relationships are causally efficacious (*Cfr.* Kim [1984] [1988]).

As I said before, what we need to guarantee that the macroproperties involved in the intentional explanations are causally efficacious is a *nomological* articulation between those properties and the mechanisms that implement them. Other formulations of Kim's seem to be too weak in this sense, because they don't require the existence of these *bridge* laws. However, paradigmatic cases of supervenience — cases such as being a liquid and causing dampness — require the setting up of those nomological relationships.

What was missing in these other definitions of supervenience is the constraint mentioned above, i.e., that a property is causally efficacious if it is a property in virtue of which the objects that instantiate it can be subsumed by laws, possibly *ceteris paribus* laws. And, in fact, it is this conjunction of positions what it seems to be behind Kim's mereological supervenience, a notion that can only be understood as a type of supervenience that necessarily implies a nomological relation (*Cfr.* Segal and Sober [1991]).

If we now put together both the definition of causality in terms of counterfactuals and this notion of mereological supervenience, we can establish a sufficient condition to guarantee the causal efficacy of any macroproperty:

If

i) there is a causal law (possibly a *ceteris paribus* law) that connects type *F* events to type *G* events and that supports counterfactuals of the kind "*G* would not have occurred if *F* hadn't taken place" and

"ii) in each case in which an *F* event causes a *G* event there exist micro-properties *m*(F) and *m*(G) such that the cause's being *m*(F) causes the effect's being *m*(G) and

iii) *F* mereologically supervenes on *m*(F) and *G* mereologically supervenes on *m*(G),

then

F is causally efficacious in the production of *Gs*" (Segal and Sober [1991], p. 10).

Once this condition has been formulated, the next step is to show that the semantic properties of mental states meets it.

5. *Cognitive and Computational Processes.*

If we had to formulate a general aim for the set of disciplines that are clustered together under the label of cognitive science, it would probably be to specify the intentional laws that govern cognitive processes and to establish the kind of mechanisms that implement those laws. So the argument that takes us to a vindication of the causal efficacy of semantic properties, i.e. the argument that shows how semantic properties meet the condition developed in the last section, falls naturally under the umbrella of cognitive science.

The core of the argument lies in the characterization of cognitive processes as computational ones. Computational processes are defined, in turn, in terms of representations. Input-representations stand for arguments in a function. Output-representations constitute the values of the computed function. A representation is thus a very special kind of physical configuration, a physical configuration that has a syntactic and a semantic reading. And the important point is that, although computer processes are only sensitive to the syntax, the machine can be designed in such a way that the production of syntactic tokens makes sense given the semantic interpretation imposed by the problems that it is meant to solve.

In the classic paradigm frame, the image of the mind as a computer implies an interpretation of intentional states in terms of states that involve symbols of a *mental* language or *Language of Thought* (Fodor [1975]). My belief that there is an apple pie in the fridge implies my being in some kind of computational relation to the *mentalese* symbols corresponding to "There is an apple pie in the fridge". The content of such an intentional state is just the content of that chain of symbols in mentalese and the fact that is a belief — instead of a desire or a doubt — is determined by the nature of its computational relation to the rest of my mental states and / or my behaviour. The symbols of that mental language possesses a combinatorial syntax and are physically implemented by the brain. Processes underlying relationships among intentional states are, in the end, physical processes.

Such an approach allows to establish causal laws that relate intentional states in virtue of their semantic properties. The nomological character of these relationships is guaranteed by the existence of those very same states

at a lower level, a physical level that constitutes the supervenience base of the properties involved at the intentional level. Those laws will probably contain *ceteris paribus* clauses since alterations or failures at the physical level may prevent the correct working of the system, but as I've already argued, that is not an important problem. The relevant point is the existence of a adequate supervenience base and, under this view, that condition is met, since the supervenience base includes all those physical properties of representations that explain their causal powers at a micro-physical level.

The answer to the question whether or not there are intentional laws of the form "Every instantiation of P causes an instantiation of Q" (possibly with *ceteris paribus* clauses) where "F" refers to a semantic property is, in the light of this approach, clearly positive. The semantic properties of mental states can be seen as causally efficacious with respect to different behaviours, although the equivalent class to which they belong is not describable in terms of features that are projectable in a physical vocabulary. Of course there must be a physical description, as there has to be a physical implementation of the properties responsible for the semantic causation, but that physical characterization is not the relevant description for the explanation of their causal efficacy (Cfr. Horgan [1989], McLaughlin [1989]).

Now, this computational approach represents a strategy according to which the ascription of causal efficacy to a certain intentional state requires some kind of *vehicle* — located at the language of thought level — that can be seen as the modular item responsible for the information embodied in that state and as the item involved in all the cognitive processes in which that state plays a role. What can we say then about the connectionist paradigm in which such wehicles seem noticeably lacking?

As I said in section 1, that question leads us to a dilemma: either we accept that connectionist models are inadequate as cognitive models or we have to adopt an eliminativist position with respect to propositional attitudes that is equivalent to deprive their content of any causal efficacy. In the rest of the paper I try to defend connectionism from *both* accusations by showing that the dilemma is, like so many, a false one. To accept connectionism as an adequate cognitive model doesn't necessarily imply an eliminativist outcome provided that we are clear about the appropriate level of analysis of such computational models.

6. *Connectionism and Levels of Description*

Connectionist models are complex networks of simple computational elements connected in parallel. Each one of these elements or units has an activation value that is established numerically as a function of both the activation values of other units in the network and the weight of the connections to those units.

The influence of a unit *a* over a unit *b* is the outcome of multiplying the activation value of unit *a* times the connection strength between *a* and *b*. Thus, if one unit has a positive activation value, its influence on the value of the adjacent unit would be positive if the connection strength is positive and negative if the connection strength is negative. In a clearly neurological reference, positive connections are called excitatory and the negative ones inhibitory.

A typical connectionist model has three set of units: input, output and the so called hidden units. The input representation is established by assiging activation values to the input units. This activation is propagated through the connections towards the hidden units until a set of activation values settles down in the output units. The computations that take place in the network to transform the input activity patterns into output activity patterns thus depend mainly on the set of connection strengths — determined according to a particular learning rule. Those connections strengths are usually considered as responsible for the *knowledge* of the system and in that sense play the same role as programs in the classical paradigm (*Cfr.* Smolensky [1988]).

Part of the interest of connectionist networks lies in their auto programming capacity, *i.e.*, in the incorporation of learning procedures through which, after a certain training period during which the network is exposed to a bombardment of input / output pairs, the network adjusts the connection strengths and establishes the functions that the hidden units have to compute.

However, this is not the most relevant feature for our discussion. The crucial point is the distributed and superpositional character of representation in connectionist models. The mechanisms responsible for the input / output relationships in these models are the units that form the network. Each unit contributes to encoding information about different representations — that is what superpositional representation means — and each representation is thus distributed across a large set of microfeatures that are subsymbolic, i.e., that are not semantically interpretable.

As the network's computations are completely determined by activity at the units level, and those units encode at the same time information of many different types, it is not possible to identify a stable and recurrent *entity* that corresponds to the classical notion of a symbol and that is handled by an independent processing system. At the same time, since the same information can be represented by networks with different units and connections, the class of networks capable of representing a fact P is nothing but a "chaotically disjunctive set" that doesn't mirror at all the natural class that would result, according to folk psychology, by considering the set of cognitive agents that have a particular belief (*cfr.* Clark [1990]).

The first step in the refutation of these arguments is a concession to the eliminativist: if the level of analysis used to explain the behaviour of connectionist models really is the units level, then the eliminativist conclusion is perfectly plausible. But, and this is the important point, a proper treatment of the connectionist paradigm — as a *cognitive model* — has to be developed at a level of description higher than that of mere numerical units and activations.

One of the first lessons learnt in cognitive science is that there are multiple levels of description with respect to a computational model. The election of one or other of those levels places strong constraints upon the type of explanations that we can give about the system behaviour. It is not thus strange that, by restricting ourselves to the units and weights level, we get unsatisfactory explanations from the point of view of a *semantically* interpreted behaviour.

However, within this paradigm there are techniques of analysis that let us go up to a higher level of description, a level of description under which we *can* identify representations and transformations that have the same role as the representations and rules of the classical systems, although with a character completely different to that of classic symbols and algorithms since those representations and *rules* are implicit and highly distributed.

One of those techniques is *cluster analysis*. The basic idea of this statistical method is to extract regularities in the activation patterns of the hidden units for each one of the input-output relations and use them to build representations that group together relations with similar patterns. This technique lets thus unify under particular semantic categories activation patterns that, at the units and connections level, have a very different structure. And most important, it lets us establish a semantic unification of the outputs that are brought about by different inputs, *i.e.*, it lets cluster together, according to their causal efficacy with respect to a particular

behaviour, what at the *units* level are just different network states(6).

The same technique of cluster analysis shows that the part of the eliminativist argument that brings into question the existence of natural classes, in the psychological sense, is unnecessarily reductionist. The fact that different networks can represent the same information despite being constituted by sets of units with different activations and weights ceases to be a problem if the analysis of those networks is carried out at a level of description where the explanatory *blocks* are not so much the units but rather the clusterings of their activation patterns. Since these patterns are the result of grouping together different physical mechanisms that cause, however, the same output, the analysis of the system at this higher level of description lets us group what looked earlier like a "chaotically disjunctive set" into a single equivalence class (*Cfr.* Clark [1990]).

Thus, the cluster analysis technique plays, with respect to the problem of establishing the suitability of connectionism as a cognitive paradigm, the same role as the analysis developed above in terms of counterfactuals and supervenience played with respect to the problem of mental causation. Both approaches offer the possibility of vindicating the causal efficacy of the semantic properties of mental / computational states despite the multiple realizability of the physical structures underlying those states. And, most important, the fact that this vindication is possible within the connectionist paradigm shows that the existence of modular items structured according to a classical syntax and semantics is not an empirical constraint essential to the success of the computational strategy.

To see this, reflect that the process of symbolic *labelling* of the activation patterns of the units in a connectionist network is a process that is carried out from *outside* the system. The patterns by themselves, are not symbolic in any sense comparable to the classic notion of symbol. They are not comparable from a semantic point of view because those patterns only

(6) If the network is complex enough, as in the case of NETtalk (Sejnowski and Rosenberg [1986]) —that turns written text into phonemes— that partitioning of the representational space will have a tree-shape hierarchical structure. In this net, the final partition between consonant and vowel is justified by the fact that the activation patterns that work as representations for each one of the consonants are more similar to each other than those that work for each one of the vowels. In turn, within the consonant group, the activation patterns corresponding to the phonetic feature *palatal*, for instance, can be grouped together as such because the similarity among them is higher than the one among the patterns corresponding to, let's say, *labial* or *nasal*.

represent sets of microfeatures and not atomic symbols. They are not comparable from a syntactic point of view because the combinatorial operations over those patterns are not governed by linguistic-type rules but by exclusively mathematical computations. This notwithstanding, their function is to characterize the information represented by the hidden units activation, the information that is causally responsible for the system's behaviour.

The characteristics of this Second Computational Metaphor (Bates and Elman [1992]) represented by the connectionist paradigm have only been sketched, but even a brief discussion like this one gives us the necessary theoretical tools to complete the other half of the aim of this paper. It shows that it is possible to justify philosophically the causal efficacy of the content of mental states *without* having to embrace a computational strategy that postulates the existence of discrete symbolic items. And, therefore, that it is possible to save connectionism from the false dilemma threatener by the eliminativist / epiphenomalist arguments.

7. *Concluding Remarks*

A variety of different problems concerning the causal efficacy of content have now been displayed. The resulting picture can be sketched as follows. There is a certain philosophical paradox about the causal powers of our intentional states. Although it is perfectly reasonable to claim that what we think is causally responsible for what we do, a more fine-grained philosophical analysis seems to show that those mental states must be pictured as epiphenomenal; a situation mirrored within the computational arena of cognitive science.

The first task was to develop an account of the causal efficacy of content that avoids this epiphenomenalist paradox. In order to do that we used the Davidsonian strategy of addressing the question about the causal role of any property P in terms of the existence of causal laws involving P (section 2). But once the scenario is set up in that way, it seems that the only strict laws possible are the micro-physical laws. The rest, the laws of the special sciences include *ceteris paribus* clauses and are therefore only approximate formulations. An anti-reductionist argument that tries to reduce this big gap between strict laws and ceteris paribus laws is developed in section 3. There I pay special attention to Fodor's arguments in Fodor [1991] and try to highlight what looks like a typical circularity mistake in his defence of the

legitimacy of those laws.

In section 4 I develop an alternative defence using as key notions the notion of counterfactual and of mereological supervenience. Once these two notions have been explained, a sufficient condition to guarantee the causal efficacy of any macroproperty is formulated. And, in order to show that semantic properties also meet that condition a computational approach is developed in section 5.

The outcome of the computational characterisation of cognitive processes used in this approach seems to imply however that we really do need to be able to isolate physical items of some kind to act as vehicles — in the supervenience base — for the higher level semantic properties. We faced a dilemma insofar as connectionist approaches then threatened to lead to eliminativism. In section 6 I showed that the correct analysis of such models needs to be conducted at a higher level (which is what techniques like cluster analysis provide). Once the correct level of description is found, the explanation of the behaviour of these models can avoid both the epiphenomenalist paradox and the eliminativist dilemma.

Washington University in St. Louis

REFERENCES

BATES, E. A. and ELMAN, J. L. [1992] "Connectionism and the Study of Change". CRL Technical Report 9202. Centre for Research in Language. University of California, San Diego.

BLOCK, N. [1991] "Can the Mind Change the World" in Boolos, G. (ed.), *Essays in Honour of Hilary Putnam*, Cambridge, Cambridge University Press.

CLARK, A. [1989] *Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing*, Cambridge, Mass., M.I.T. Press.

CLARK, A. [1990] "Connectionist Minds", *Proceedings of the Aristotelian Society*, vol. XC, pp. 83-102.

CLARK, A. [1991] "Radical Ascent", *Proceedings of the Aristotelian Society*, vol. sup. LXV, pp. 211-227.

CLARK, A. [1993] *Associataive Engines. Connectionism, Concepts and Representational Change*, Cambridge, Mass., M.I.T. Press.

CLARK, A. and KARMILOFF-SMITH, A. [forthcoming] "The Cognizer's Innards: a Psychological and Philosophical Perspective on the Development of Thought". *Mind & Language*.

CHURCHLAND, P.[1986] *Neurophilosophy*, Cambridge, Mass., M.I.T. Press.

CHURCHLAND, P. [1989] *The Neurocomputational Perspective*, Cambridge, Mass., M.I.T. Press.

DAVIES, M. [1991] "Concepts, Connectionism and the Language of Thought" in Ramsey, W., Stich, S. and Rumelhart, D. [1991], pp. 229-257.

DRETSKE, F. [1988] *Explaining Behaviour: Reasons in a World of Causes*, Cambridge, Mass., M.I.T. Press.

FODOR, J. [1975] *The Language of Thought*, New York, Thomas Y. Crowell.

FODOR, J.[1987] *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, Cambridge, Mass., M.I.T. Press.

FODOR, J.[1989] "Making Mind Matter More", *Philosophical Topics*, 17, 1, 59-79.

FODOR, J. [1991] "You Can Fool Some of the People all of the Time, Everything Else Being Equal; Hedged Laws and Psychological Explanations", *Mind*, vol. C (1), pp. 19-34.

FODOR, J. and PYLYSHYN, Z. [1988] "Connetionism and Cognitive Architecture", *Cognition*, 28, 3-71.

HORGAN, T. [1989] "Mental Quasation", *Philosophical Perspectives*, 3, pp. 47-74.

JACKSON, F. and PETIT, P. [1988] "Functionalism and Broad Content", *Mind*, 97 (387), pp. 381-400.

JACKSON, F. and PETIT, P. [1992] "Causation in the Philosophy of Mind" in Clark, A. and Millican, P. (eds.) *Proceedings of the 1991 Turing Colloquium*, Oxford, Oxford University Press, forthcoming.

KIM, J. [1984] "Epiphenomenal and Supervenient Causation" in French, P. et. al. (eds.), *Midwest Studies in Philosophy, IX*, University of Minnesota Press, Minneapolis, 1984.

KIM, J. [1988] "Supervenience for Multiple Domains", *Philosophical Topics*, 16 (1), pp. 129-150.

LEPORE, E. and LOEWER, B. [1987] "Mind Matters", *Journal of Philosophy*, 84, 11, pp. 630-642.

LEWIS, D. [1986] *Philosophical Papers*, vol. 2, Oxford, Oxford University Press.

McCLELLAND, J., RUMELHART, D. and the PDP Research Group [1986] *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1 y 2, Cambridge, Mass., M.I.T. Press.

McLAUGHLIN, B. [1989] "Type Epiphenomenalism, Type Dualism and the Causal Priority of the Physical", *Philosophical Perspectives*, 3, pp. 109-135.

PINKER, S. and PRINCE, A. [1988] "On Language and Connectionism. Analysis of a Parallel Distributed Processing", *Cognition*, 28, pp. 73-193.

RAMSEY, W., STICH, S. and GARON, J. [1991], "Connectionism, Eliminativism and the Future of Folk Psychology" en Ramsey, W., Stich, S. and Rumelhart, D. [1991], pp. 199-228.

RAMSEY, W., STICH, S. and RUMELHART, D. (eds.) [1991] *Philosophy and Connectionist Theory*, London, Lawrence Erlbaum.

SCHIFFER, S. [1991] "Ceteris Paribus Laws", *Mind*, vol. C (1), pp. 1-17.

SEGAL, G. and SOBER, E. [1991] "The Causal Efficacy of Content", *Philosophical Studies*, 63, 1-30.

SEJNOWSKI, T. and ROSENBERG, C. [1986] NETtalk: *A parallel network that learns to read aloud*, John Hopkins University, Technical Report JHU/EEC-86/01.

SMOLENSKY, P. [1987] "Connectionist AI, and the Brain", *Artificial Intelligence Review*, 1, pp. 95-109.

SMOLENSKY, P. [1988] "On the Proper Treatment of Connectionism", *Behavioural and Brain Sciences*, 11, 1-74

STICH, S. [1983] *From Folk Psychology to Cognitive Science*, Cambridge, Mass., M.I.T. Press.

VAN FRAASSEN, B. [1980] *The Scientific Image*, Oxford, Clarendon Press.