Abstracta 2 : 1 pp. 3 – 23, 2005

MENTAL EXPLICITNESS: THE CASE OF REPRESENTATIONAL CONTENTS

Pierre Steiner

Abstract

This paper aims at answering the question "When is informational content explicitly represented in a (human) cognitive system?". I first distinguish the explicitness this question is about from other kinds of explicitness that are currently investigated in philosophy of mind, and situate the components of the question within the various conceptual frameworks that are used to study mental representations. I then present and criticize, on conceptual and empirical grounds, two basic ways of answering the question, the first one coming from the classical computational theory of mind, the latter one issued from a procedural conception of informational contents. I then argue for a new answer to the initial question, an answer that retains some valuable insights of the criticized theories, but which underlines the importance and the interpretational source of the distinctiveness proper to explicitly represented contents.

Introduction

Amongst the numerous terms that are widely used but not so much accurately understood in philosophy of mind and in cognitive science we can find the adjectives "explicit", "implicit", and "tacit". We can apply them to psychological attitudes (as when we speak of explicit knowledge, implicit learning, tacit belief), but also to informational contents. What must be noted is the fact there is not a mutual determination between these properties of attitudes and these properties of contents (while the fact an attitude is such-and-such might depend on other properties exhibited by its content): for example, what is explicitly represented within one's mind is not necessarily the object of an explicit psychological attitude (knowing, believing, attending...). Consider phenomena like change blindness, *priming*, implicit learning, and - more debatable - blindsight cases: according to the conditions I'll set later on, the unconscious contents that are here processed and that can influence one's behavior are probably explicitly represented (at least they do have neural correlates[1]), while implicitly or tacitly known (they occur at the subpersonal level). Or, more elementary, think about all the information processed by the visual system: most of them can *never* be made conscious or be objects of explicit knowledge. Representations *in* the system (explicitly tokened) are not necessarily representations *for* the system (explicitly known).

[1] With a caveat for change blindness.

The criteria allowing us to classify a psychological attitude as explicit-implicit-tacit (criteria that may consist of conscious access, verbal report, voluntary control towards the content of the state[2]) are not the same as the ones allowing us to classify an informational content as explicitly represented. This set of considerations – that some will consider as mere platitudes - should be reminded every time one begins to talk about "implicit", "explicit" or "tacit" knowledge and representations just because unconsciousness has been met around the corner[3]. Indeed, concerning mental representations, (un)consciousness is neither necessary nor sufficient for a relevant definition and study of *all* the explicit-implicit-tacit properties they can display.

In what follows, I shall focus on explicitness of content. The question I'll try to answer is the following: when is an informational content[4] explicitly represented in a human cognitive system? Beside its terminological relevance, I consider it as crucial for every representationalist theory of the mind. There are many ways one can be a representationalist and thus consider the conditions under which a mental representation may be said to be present and causally efficacious within a certain kind of system (be it central and abstract, embodied, extended, coupled...). Whatever the theory you defend, if you want your theory to make a flexible use of the notion of representation, you'd better accept that information may come, be represented and causally efficacious under different kinds of storage.

I shall present and criticize two ways of answering to the question: the classical answer (defended by Fodor and Pylyshyn) and the procedural answer (Clark, Clapin, Kirsh). I'll then propose my own answer, closer to the classical one than to the procedural one, but with some decisive changes. But before all of this, some things are worth reminding in order to settle in an accurate way the issue of this paper.

---

[2] Cf. Dienes and Perner (1999).

[3] I take it that lots of philosophical controversies about *tacit beliefs* (Lycan 1986) would gain significant progress if this basic distinction were respected.

[4] Semantically and grammatically, that content may be very poor. It doesn't need to be propositional/conceptual. Issues about the semantic structures of contents (as in predicateness, hypothetical inferences,…) are here left aside. I just focus here on how contents and their structures might be implemented (I also put another question aside: how many contents do a representational system need? What features of the world and of the body need to be exactly represented and mirrored in our cognitive systems? Less than we may think, I suppose).

I. Physical presence and varieties of being stored

In order to understand (mental) representations, basic distinctions are generally made between, at least, *vehicles* (representing structures), *referents* (represented structures), *contents* (what is carried, stored, tokened, in a vehicle) and *media*. Roughly speaking, the vehicle is the particular material entity that is about something (the referent) by carrying some content (I assume here that content is an abstract entity that is instantiated in the physical properties of the vehicle – but there is often a theoretically unbridgeable gap between the content we, as interpreters or designers, attribute to the vehicle and the content really instantiated by the physical properties of the vehicle[5]). Mental vehicles take place in the brain, by being patterns of neurons or synaptic connections. By *medium* is meant the material the vehicle is realized in. Anything material can serve as a medium for tokening one vehicle/content: ink, sand, silicon chips, sound waves, etc. In the case of mental representations, it is a safe bet to say the medium is a neural one. A classical way of classifying vehicles relies on the variety of relations that may exist between them and their referents, viz, on the possible relations that may ground the fact they exhibit some aboutness property towards some states of affairs. The basic distinction, here, is between propositional and analogue modes of representation (strictly speaking, this is more a difference than a distinction: many representational phenomenon can be made of both these modes, especially in the case of mental representations). But once one of these modes of relation between the vehicle and its referent is defined as being the basis of the particular representing relation (the very basic criterion being first-order resemblance, but also the forms of processing used by the representational system) come other sub-categories for classifying vehicles. In the case of propositional modes of representation, for example, as the relation between the vehicle and its referent is not *intrinsic* (the features of the represented structure the representing structure is sensitive to and preserves are arbitrarily selected and represented), it can be caused by various kinds of semantic grounds: convention, second-order resemblance, inferential role, causal/nomic relation, functional/teleological stories to name but a few. Let's assume these various relations and modes of representation can define the relations that may exist between the vehicle and the referent, and may thus answer the question: *Which kind of relation makes the representing be about the represented*?

[5] By the way, the descriptive dimension of content often depends on what the content user *can* and *needs to* do with(in) its environment.

There is also another way of classifying vehicles, this time related to the way they code their contents. This classification defines the possibilities we may refer to when we try to answer the question: *How is informational content represented in the vehicle?* If we consider the possibility that mental representations base themselves on non-analogue modes of representation, there are at least two existing ways (and an intermediate one) for vehicles to carry content; these two ways are often too quickly associated with the two dominant paradigms in cognitive science, namely classical computationalism and connectionism.

1) Local storage: every vehicle stores (has the function of storing) only one content. Conversely, any content is represented by only one vehicle. Each vehicle has a discrete semantic interpretation, which is taken to be transcontextual (pace classical computationalism). Vehicles are symbolic units that can combine each other in order to form complex vehicles (symbols chunks) carrying complex contents. As symbols combination is necessarily based on compositional laws, the meaning of the symbols string is only determined by the meaning of its parts. Classical cognitive science is implicitly based on this kind of localist implementationism.

2) Distributed storage: A friend of the local storage thesis interested in neurosciences could be struck by the fact that, in the neural reality, what he takes to be symbols strings supervene on neural units which are not straightforward symbolic, but rather subsymbolic: neural atoms that, on their own, do not represent anything. From this, the local coding thesis should be amended: the (atomic) vehicle of (atomic) content supervenes on a pattern of subsymbolic units (neurons). On a macroscopic view, coding may thus seem to be local, while on a microscopic view it is distributed on moving patterns of neurons. It is distributed simply because it is *extended*. But, as long as (a) this distributed and supervening coding only concerns the implementation of symbolic atoms and strings and do not affect their transcontextual meaning, which is vehicle-independent and easily movable; (b) the relations between both subsymbolic units and symbolic atoms and between symbolic atoms and symbolic molecules still obeys to compositional laws and (c) distributed coding doesn't equate with overlapping and superpositional coding (the basic subsymbolic units only take part at *t* to the implementation of one vehicle), this fact is compatible with the basics of localist symbolic coding – though its acceptance turns to be necessary if the friend of classical symbolicism doesn't want to go against some very basic neural evidences (there are no grandmother neurons, for example).

3) Superpositional storage: as soon as distributed coding turns to be considered as superpositional coding, as influencing the semantic contents of vehicles (content becoming context-sensitive, viz, sensitive to the properties of its subsymbolic units that do take part in other codings), as being the *form* symbolic *complex* contents necessarily have and as being inherently non compositional *in a spatially concatenative sense* (the spatial parts of the distributed representation being not its semantic constituents), things and times are changing – for better or for worse. The local, but also compositional, features of classical symbolic coding are thrown overboard, so that we cannot see representational vehicles as symbols, or at least as belonging to any symbol *system* in the classical sense. Representations are no more discrete, stable and localized. Moreover, as atoms or as complex states, they merely supervene on sets of units that interact without following *any* compositional rules. The objects of computational processes (neurons and their firing rates) are no more symbols and cannot have any semantic interpretation on their own. Superpositional storing *starts* from distributed coding, but goes well beyond extendedness. It denies that subsymbolic units can only be used to represent one content, at *t* or during their (short) life: at *t*, every subsymbolic unit can be a part of an important number of representational patterns, which thus partially overlap. Following superpositional coding, two occurrent contents, *c*1 and *c*2, can even be represented by all the same physical resources (subsymbolic units) at the same time. They might have the same vehicle, which thus has two representational functions. A caveat is that this kind of total superpositional (totally overlapping) coding is mainly used for the storage of abeyant contents, not occurrent ones: for example, at *t*, a synaptic structure stores different pieces of knowledge that are not used by the system at *t*. This is classically called *connection weight representation*, as contrasted with local or (weakly) distributed representation. It is only by being superpositional (and thus denying spatial compositionality, stability of meaning, intrinsicness of formal properties and semantic atomism) that distributed coding of content can be theoretically relevant and thus stand against classical symbolicist coding, be it (macroscopically) local or (microscopically) distributed (see van Gelder 1990 for such an adamant view and helpful terminological points).

In their influential paper on implicit and explicit knowledge, Dienes and Perner hold that a content of knowledge is explicitly represented in a representational system when “there is an internal state whose function is to indicate the content of the knowledge”(1999:737). From the

last couple of distinctions I mentioned above, this definition clearly sounds insufficient: can this internal state be distributed on subsymbolic states? Or even be overlapping with other states? What about an internal state superpositionally storing *two* contents? Connectionist weights are also representations, with the function of indicating relations in the world. Would they also count as explicitly storing pieces of information? If explicit representing is simply representing, what is therefore implicitness? What's the indication relation between the state and the content? Does the structure of the state have to reflect the shape and the individual character of the linguistic expression of content? As I see them, the issues concerning explicit representation of content are related to the forms of storage the representational systems use. That is, the question "when is information explicitly represented?" is related to the question "how can information be represented/encoded?". But while you should better answer the latter before thinking about the former (and this is the reason of my recent detour), you should not equate explicit representation with one of these forms of coding. As we'll see, explicit presence is compatible with various forms of coding, and is defined by properties that sometimes do not appear in the characterization of these various forms. Also, some properties that are classically seen as being essential in the study of mental representations (syntactic shape, atomism,...) might turn to be irrelevant when we come to deal with the elemental question we shall consider here.

Time to start now. I'll begin with the classical computational answer to the question and will point to its inadequacy; I'll then evoke the procedural answer and the confusions it relies upon, before proposing my own answer.

II. Explicit content, the classical view

According to the classical computational theory of the mind, an informational content is explicitly represented in a system if there is a distinctive physical unit that encodes it by means of its particular syntactic features. In several papers, Fodor and Pylyshyn separately seemed to defend such a point of view. In an essay on Chomsky's regular representationalism, Pylyshyn defines explicitly represented prescriptive contents (rules) as follows:

> The rules are explicitly encoded if and only if there exists some mapping from the rules as inscribed in some canonical notation and the physical states of the system (Pylyshyn 1991:238).

A caveat must be made, related to the partial character of this definition: while they defend the view that every object of computation has to be explicitly represented, classicists do not hold that the computational rules Pylyshyn alludes to always have to be explicitly represented, for an obvious reason: Achilles and the tortoise's regress. Computational processes consist in rule-following, but, ultimately, in order for the system to act and for the rules to be understood, rules have to be instantiated in hardwired processes (Fodor 1987:23; Fodor and Pylyshyn 1988). This definition exemplifies a common intuition on the way explicitness is classically seen: as *physical presence*, here in *states* of the system. What is explicit is what is present, here - in the mind -, as physically and distinctively present (easily seen). But the quote doesn't say anything about the grounds on which way the mapping relation is defined between the notation and the physical states (isomorphism? homomorphism? functional characterization?) and, especially, about the forms under which contents are carried by the states (is superpositional presence a case of explicit presence?). Obviously, anyone aware of Pylyshyn's credences knows that, in this context, physical presence is symbolic presence. That is, the physical states of the system are physical objects that do possess a distinct semantic value and individual formal properties (by virtue of which they are objects of computational processes). There is homomorphism between *canonical* intentional states and computational symbolic states. Semantic properties are preserved through the mapping because of the necessary formal similarity between the notation and the states (syntax mirrors semantics). Formalism here comes with localism.

As usual, Fodor's views on the question are more clear-cut:

> Explicit representation is, I suppose, representation by intrinsic properties of a situation. (In the standard case, the 'situation' will involve the tokening of a symbol, and it will be intrinsic properties of the symbol that explicitly represent information that the situation contains). […] *Being explicitly encoded means being encoded by 'syntactical' (if you prefer, by intrinsic) features of mental representations* (Fodor 1987b:67 ; emphasis mine)

In order to be explicitly represented, an informational content has to be physically present in the system, in virtue of a vehicle that has to possess an *intrinsic syntactic* structure and a single semantic value. The criterion of physical presence always plays a role in the definition of explicitly represented content, but it is now made more accurate and comes with a fundamental property, namely syntax, that will allow the physical state to have a spatially individuated

character and to play a systematic causal role. Syntactic properties are essential and intrinsic to the physical state – the syntax of a complex physical state is totally determined by the way its constituents parts are related (Fodor 1994: 4), so that two states with the same syntactic features are functionally identical. In Fodor's theory, "formal", "syntactic" and "non semantic" are rough synonyms (1982:100)[6]. It is very often assumed that, by being syntactically structured, representational vehicles should therefore be *physically* discrete and visible to inspection. But we must bear in mind that the syntactic structure is not necessarily physical: pace Fodor, syntax is not exactly shape, it is an higher-order physical property that might eventually, in some cases, reduce to shape (1987:19), so that, on purely logical grounds, the discreteness Fodor requires could be only functional, and not physical. But the functionally-defined structural properties of symbols, in the classical model, are anyway very often (dangerously!) supposed to correspond to *real* physical structures in the brain (Fodor and Pylyshyn 1988:13). The compositional and linguistic structures of mental symbols parallel the structures and formal properties of neural events and processes. The fact the symbol possesses its causal role by virtue of its syntax and the fact syntax is exemplified through shape implies that syntactic properties and physical properties mutually constrain each other, so that whatever the level of analysis, explicitness of content, in Fodor's theory, is necessarily related to basic syntactic physical presence (symbolic presence). But it seems unlikely, as I'll show next, that informational content can be physically encoded that way.

We must also note that Fodor's above definition has an obvious implication:

> The formality condition and the explicitness condition are going to have to stick together if they're going to stick at all (1987b:75).

The formality condition (mental states are causally efficacious in virtue of their form) is thus valid if and only if mental representations are explicitly represented (viz, having an intrinsic syntactic shape); and, conversely, mental representations have to be explicitly represented in order to take part to causal and structured trains of thoughts. This equation allows Fodor to recast his "no computation without representation" motto into: "No intentional causation without explicit representation"(1987: 24-25)

[6] See Devitt (1991) for helpful remarks on the idiosyncratic Fodorian use of "syntax", "form" and "shape".

This motto only applies to so-called *core cases* of mental representation, that is, to cases where the mental representation is supposed to be occurrent and to play a causal role in the system: in these cases, the intentional content has to be explicitly represented. This applies to propositional attitudes that are causally involved in the causation of behavior, and not in its mere description (as dispositional states, for example) (Fodor 1987:22). Each time a thought is supposed to play a causal role in our mental life, it is by having its content explicitly tokened within the system. Aside from the fact it is real, intentional causation necessarily comes with explicit representations, the only form of representation that can honour the formality condition. Sure, intentional causation, according to Fodor, is also accompanied with other forms of representations and processes that don't always need to be explicitly represented: rules, or, ultimately, hardwired processes. These processes take explicit representations as objects. But Fodor's ontology, in that domain, remains quite conservative: representations are either explicit or simply dispositional, and are *data structures*, viz, objects of explicit rules, or, ultimately, of "implicit" rules (hardwired processes).

From these two features of explicit representations (they have to be syntactically structured and they are the necessary form under which a content has to be tokened in order to be causally efficacious) two criticisms may result:

(1) According to Fodor, because of its intrinsic syntactic properties, the symbolic vehicle necessarily possesses the following properties: it is stable, static, movable, discrete, while the content it conveys would be context-independent. These properties of mental vehicles might be proved to be empirically not very plausible: facing the distributed and superpositional representational properties of brain *processes*, is it safe and necessary to see explicitness as belonging to well-defined symbolic *states*? Local, determinate and complete codings of distal properties, or even intrinsically structured distributed coding, turn out to be more like theoretical posits than neural realities[7]. Indeed, possibilities of local coding of content in one single neuron (the grand-mother neuron), or in stable groups of neurons that would only be devoted to the representation of only one content are clearly dismissed in contemporary neurosciences (see e.g. the way Crick & Koch (1995) consider explicitness). The physical presence of semantic contents in the brain is not like the physical presence of these same items on a sheet of paper. The friend

[7] See, e.g., P.M. and P.S. Churchland (1998:39-44), P.S. Churchland (2002:chap.7).

of classicism will surely argue that this kind of neural-based criticism against the syntactic and symbolic features of classical computational models is deeply irrelevant, but also problematic. It is irrelevant because (so the song goes) "it is more about the physical realization of cognitive states than about what classical theories aim at studying: the systematic and productive workings of cognitive states and processes". It would be problematic, because, to say the least, cognitive theories intending to be too much brain-friendly and rejecting symbolicist basis would show notable problems in order to account for some essential features of thinking. I'll answer to this second objection while developing my second criticism. For now, let me just say that the relevance objection just begs the question. If there's one dimension of mental representations that might be related to physical presence and implementation, it must be their (possible) explicit character! Wishing to define explicit representational presence by paying no attention to the physical reality of information storage is not very coherent. Or, at least, it is not defining what explicit presence actually is. There's a huge difference between thinking about the physical form and realization of mental representations only from constraints that are set by your theoretical requirements (as Fodor does) and considering the neural plausibility of your constraints by paying attention to the physical medium mental representations take place in. The argument from neural plausibility doesn't primarily aim at criticizing Fodor and Pylyshyn's whole definition. It rather aims at showing that defining explicitness in terms of *syntactic* physical presence *might* prove to be wrong. And – this is my main point here – the fact is we do not even need to mention the (possible) possession of syntactic properties in order to define the criteria allowing us to answer the question "when is information explicitly represented?", the main criterion of explicit presence being merely physical presence (and some kind of distinct coding). Even if syntactic properties did exist, they might not be determinant to help setting what it is for a content to be explicitly represented, as they are not related (and cannot be found!) to this basic level of implementation and they might not have the properties Fodor thinks they have (intrinsic, transcontextual). The importance of syntactic properties may only appear at a level of analysis where issues concerning explicitness have already been settled. Moreover, syntactic properties might turn to be more holistic and contextual than intrinsic and local, as for example Stich sees them:

> Mental state tokens are brain state tokens. But the properties in virtue of which mental state tokens are classified into syntactic categories are not intrinsic features of those brain states; they are not features which depend exclusively on the shape or form or "brute physical" properties of the states. Rather, the syntactic properties of mental states are relational or functional properties – they are properties that certain states of the brain have in virtue of the way in which they causally interact with various other states of the system (Stich 1991 : 244).

From this point of view, I suppose connectionists would be ready to embrace the view that mental representations do have syntactic properties. I take it that when Fodor says, in a footnote, that "*any* nomic property of symbol tokens (...) – any property in virtue of the possession of which they satisfy causal laws"[8]) could realize a syntactic property, he's closer to the truth than he may think. Why, after all, couldn't nomic properties of brain states be holistic?

I now turn to my second criticism:

(2) The requirement "No intentional causation without explicit representation" is not very ontologically economic and explanatory useful. It implies you have to posit lots of representational structures you can *reckon* (they are functionally discrete and neat) in order to explain a basic cognitive action[9] and it excludes various forms of mental causation that do not only rely on explicit representations (exploiting environmental constants, background knowledge on relevance,...). By considering ways neural networks process information we can see how mental contents (like general knowledge, tacit beliefs) can influence the processing and the transitions from representational patterns to representational patterns without being themselves tokened (and syntactically structured) and considered by the system[10]. While it is dubious there could be intentional causation without at least *one* explicit representation, it is clear that, following neural networks, mental content can occur in intentional causation without being explicitly represented. With only explicit representations, implicit rules and hardwired processes, intentional causation doesn't go a long way. For example, the tokening and the role of an explicit representational content in a system often causally depend on representational contents stored in synaptic weights (background knowledge) (see e.g. Churchland and Sejnowski (1992:168)). The representational and working space of a cognitive system isn't limited to explicit tokens (but also,

[8] 1987:156 n.5 ; Fodor's emphasis

[9] For example, Fodor and Pylyshyn (1988:57): "conventional architectures requires that there be distinct symbolic expressions for each state of affairs that it can represent".

[10] To be fair, it seems that Fodor almost foresees this possibility when he for example remarks that explicitly tokened thoughts "might produce structural changes that support being disposed" to a style of thinking, before being erased (1991:292).

on another extent, to an *inner* space). Moreover, accepting that thought contents may be stored, but also causally efficacious and updated without being explicitly tokened (in the classical sense) could set aside some basis of the (in)famous *frame problem* (Bickhard and Terveen 1995).

Once again, by this criticism, I do not call the whole of the classical paradigm in question. I rather wish to insist on the fact that some of its main problems are bred by its (unrealistic) requirements (the only representations that get involved in mental causation have to be explicit) and could thus be avoided if only attention were paid to the existence of other forms of representation, like the ones we can find in neural networks. Put otherwise, I do not see classical computationalism and connectionism as being totally opposite or even *rival* paradigms. Time will tell if connectionist architectures are really unable to account for the systematicity and the productivity of thinking without positing (spatially concatenative) compositional states and a language of thought. What neural networks theories offer us about representation storage is anyway very relevant for any real cognitive theory. Neural stuff doesn't merely implement information processing; it also determines and produces parts of the representational and systematic resources by which information processing can occur. The representational resources of connectionist systems (lots of them being non explicit, in the classical sense) can be used to make better accounts of mental causation, if we accept that syntactic form, in Fodor's sense, is not a fundamental and intrinsic property of all representational states.

The criticisms I made to this classical conception of what it is for a mental content to be explicitly represented were more aimed at both refining it with new resources and alleviate it of unnecessary features than at rejecting its leading idea. Physical *distinct* presence indeed turns to be a necessary condition for a content to be explicitly represented. But *distinctiveness*, as I'll show, is no necessarily physical. The proponents of the procedural conception of explicit contents have nevertheless criticized this last idea in recent years. Their criticisms are very close to the ones I used earlier in order to criticize the lack of neural plausibility of the classical conception; but the position they instead propose turns to be incoherent, as I'll try to show now.

III. Explicit content, the procedural view

Underlining the limitations and dubious symbolicist assumptions of the classical view, several authors, such as H.Clapin (1999), A.Clark (1992, 1993:chap.6) and D.Kirsh (1990, 2003) put forward an alternative theory in order to define the explicitness of mental contents. This theory is

dubbed *procedural*, for it bases itself on a procedural definition of explicitness: according to it, an informational content is explicitly represented in a cognitive system if it is easily accessible and usable for that system, in an inferentially integrated way (the system has to be able to use the content in various circumstances and in interaction with other states (see Evans's Generality Constraint)). Ease of use and multiple exploitability/deployability are jointly necessary and sufficient to define what it is for content to be explicitly represented. In his 1990 paper "When is Information Explicitly Represented?", Kirsh points to the fact we should not let ourselves too quickly abused by sentential and linguistic requirements when we investigate mental explicitness. Pace Kirsh, linguistic explicitness (found in words, sentences,...) is grounded upon four properties it would be dubious to *directly* attribute to mental representations (Kirsh 1990:342):

1) Localization: words are visible and have a definite localization
2) Movability: words retain their meaning across contexts (notwithstanding indexicals)
3) Determinate meaning: words have a definite meaning
4) Availability: the meaning of words and sentences is directly accessible (without interpretation or translation) to the people understanding the language.

Three of these four linguistic properties can be found in the symbolic paradigm of cognition (the property of availability for the system or subject cannot indeed be found in modular representations, for example). According to the procedural theorists (especially Clark and Clapin), as distributed and superpositional storage make (1), (2) and (3) impossible for mental representations, we should focus on (4) to define what is mental explicitness. Clark and Clapin deny the priority of physical presence over accessibility in order to define the conditions under which content is explicitly represented. The accessibility of the informational content, but also the fact it is usable in various ways are collectively sufficient in order to define the conditions stipulating when a content is explicitly represented. For example, knowledge superpositionally stored in synaptic connections at *t* can be said to be explicitly represented in the system, for it can easily be unfolded by the cognitive system at *t*, while it does not have a distinct and structured physical vehicle: pace Clark, "the implicit-explicit continuum is (...) better viewed as a two-space whose dimensions are, first, ease of usability of information and, second, variety of *modes of use*" (1992:198). According to Clapin, "determining whether certain information is represented

explicitly or inexplicitly depends on the availability of that information to various processes in the system"(1999:151). The explicit presence of the representation is not structural, isolable or simply visible; it is essentially processor-relative, that is, only related to the *abilities* of the system[11]. Kirsh still seems to consider distinct physical presence as necessary, but not as sufficient for defining explicitness; availability must indeed be included, and has a prior role to play as criterion[12]: "Explicitness really concerns how quickly information can be accessed, retrieved, or in some other manner put to use. It has more to do with what is present in a process sense, than with what is present in a structural sense"(1990 : 361).

The procedural view relies on criteria totally opposite to those put forward by classicism. Indeed, for the latter, distinct (syntactic) physical presence is the key, while accessibility for the subject or the system is irrelevant. According to the former, accessibility to the system is what matters, and not distinct or localized physical presence (for Kirsh, it seems both are necessary). Thus, a piece of semantic knowledge which is not used somehow by the system at *t* and which is stored in connections weights will said to be nonexplicitly represented for the classicists, while explicitly represented for the friend of the procedural conception. One can agree with the empirical reasons leading people like Clark, Clapin or Kirsh to criticize the classical theory and its sententialist, formalist and objectivist assumptions. I myself used the same kind of neural-based arguments in order to criticize the plausibility of some formal conditions it defended. Still, I think proceduralists go too far in the inferences they draw from their criticisms, and thus end with a theory – the procedural one – I find quite dubious and more problematic than the classical one, as it seems to be both too liberal and restrictive for defining mental explicitness, and this for at least three distinct reasons:

(1) In order to define the explicit character of an informational content, this theory relies too much on properties belonging to or produced by the attitudes of the subject or of the system (like easiness of access and variety of uses), and not enough on properties more specific to the

[11] This kind of theory might be seen as one of the basis of Clark's (and Chalmer's) extended mind theory: the mind is not inner stuff, but is made of all the *external* information we may use and *environmental* devices we may exploit in our cognitive life. While I am sympathetic with this *transcranialism*, I refuse to consider explicit *brain* representational structures in this procedural way.

[12] He thus turns the four linguistic conditions under which a content is explicitly represented into four other inferential-procedural conditions I shall just enumerate: 1) vehicles must be easily separable from each other ; 2) it is trivial to identify the semantic and syntactic identity of the symbol ; 3) symbols must be readable in constant time and able to fall in the attention span of an operator ; 4) the information a symbol explicitly encodes is related to the set of processes it activates in constant time.

informational vehicle. The theory therefore mistakes criteria related to the attitudes and abilities of the system with criteria proper to the representational vehicle. Actually, it doesn't offer us a theory on the physical presence of representational content, but a theory of cognitive accessibility for representations. This confusion brings a lack of fineness of grain that has at least the two following counterintuitive consequences:

(2) The criteria put forward by the procedural view rule out the possibility that modular representational structures, or inferentially insulated informational states, be explicitly represented in the system at *t*, while they seem to be causally efficacious and seem to be physically present in the system, with a distinct format. Sure, they are unconscious, viz implicitly/tacitly known, but that doesn't mean they are not explicitly represented. If it is true and important that some explicit representational states display easiness of access and are variously deployable, it is not always so, so that this establishment is not very interesting when we try to answer the question "when is information explicitly represented?" (although it might be for the different question "when is information explicitly known by the system?").

(3) The procedural view blurs a basic metaphysical distinction between occurrent and dispositional (or "abeyant") representational states. Explicitness – for content, but also for attitude - is tied to *presence*, to *occurrence*. There is a basic difference between being *accessible* in *t* and being *accessed* in *t*. The procedural view is unable to account for the physical distinction between dispositional and occurrent states, for it says both of them are explicitly represented. This distinction is important, ontologically and scientifically: brain processes involves *occurrent* states (but not necessarily explicitly tokened in Fodor's sense), not dispositional ones. Every theory of mental explicitness should account for this distinction; if it doesn't include it somehow (without perfectly embracing it, of course), it wont' be very relevant.

My criticisms do not aim at proving the procedural view is inherently wrong, but rather at showing the taxinomy it defends is counterintuitive and inadequate if it wants to cope with the diversity of cognitive states. While it rightly emphasises the importance of the use of representations in a cognitive system, the procedural theory fails to see that variety of uses fundamentally requires varieties of storage (that is, varieties of physical presence).

IV. An alternative proposal

I now wish to propound another view allowing us to define what it is for an informational content to be explicitly represented in a cognitive system, without adhering to the classical and procedural views. By this definition, I hold you can still buy explicit representation without embracing Fodor's unrealistic requirements and the procedural conditions. The definition is the following:

> An informational content *p* is explicitly represented in *S* at *t* ≡ ∃ a distinct physical structure *V* whose occurrence has the function of carrying *p* in *S* at *t*.

The main point is: *the distinctiveness of the physical structure is not necessarily physical*. That is, by supervening on a set of neural units that can take part to the activation of other vehicles, the physical structure doesn't need to have accurate and stable spatial boundaries. The only distinctiveness that is required is the one that will allow us to give a single semantic value to the vehicle. Semantic distinctiveness is not reflected in a hypothetical physical discreteness. Semantic discreteness depends more on functional neatness than on physical distinctiveness. Actually, the attributed semantic value is what allows us to *recognize* the vehicle amongst all the sets of firing neurons; as the semantic value is attributed to the vehicle that shall best mirror (or approximate) the functional role implied by the semantic content. At *t*, the identified vehicle has to have a single semantic character, so that it may only be said to carry one semantic content (it is supposed its tokening has the *function* of carrying one semantic content) – but at *t*, various parts of the vehicle may take part to the occurrence of others vehicles and therefore code various parts of other semantic contents carried by other firing vehicles - but the whole of all the firing neurons cannot take part to the activation of just another vehicle. While one unit can take part to the explicit representation of many contents, a set of units or physical states cannot represent more than one content in order to encode it explicitly. Explicit representation of content can be realized on a distributed mode; it only excludes *total* overlapping and (thus) superpositional storage from being explicit representing. By total overlapping (superposition) I mean the fact *all* the firing parts of the vehicle *V1* are also the complete firing parts of another occurrent vehicle *V2*, so that *V1* is not physically distinguishable from *V2*. If this occurs, then a single semantic value cannot

be attributed to the vehicle, so that it cannot be said to explicitly encode content. This we can often find in the case of contents encoded in connection weights (one weight implicitly stores numerous contents, usually general features of the world). But note that some parts of *V1* can overlap with some parts of *V2*, and other parts of *V1* can overlap with parts of *V3*. Even if *all* the parts of *V1* were to overlap with parts of many other vehicles that would even be enough to semantically individuate *V1* from the other vehicles, if we pay attention to the differences between the activation times of the vehicles, to which other vehicles they project, and how they are connected.

Three remarks are still required in order to make this proposal more accurate:

(1) No syntactically and intrinsically structured vehicles are required in order for a content to be explicitly represented (this is against the classical view). Moreover, depending on the complex character of content you want to attribute to a part of the system, the physical structure can be discontinuous: the representation of an event or object can consist of various representations of its aspects, distributed over parts of the system. The vehicle of some higher-order content can thus be divided.

(2) A vehicle that, at *t*, superpositionally codes various contents cannot be said to explicitly represent a particular content. But while something superpositionally stored is, by definition, not explicitly represented, it can nevertheless be causally efficacious: *causal role does not require explicit representation* (see the role of connection weight representation in neural networks);

(3) Enduring stability (temporal or local), distinctiveness and force are not necessary properties of explicit representations. Still, stability (the fact that the neurons of the vehicle fire at a constant rate) seems to be necessary for the content to be conscious, but not sufficient[13] (we are not aware of the stable patterns of activation over the light receptors of our retinas, although it can be said they do code some non-conceptual information).

V. Conclusion

More than twenty years ago (in a paper reprinted in his *Intentional Stance*), Dennett defined as follows explicitly represented information:

[13] This is against O'Brien and Opie (1999).

> Information is represented *explicitly* in a system if and only if there actually exists in the functionally relevant place in the system a physically structured object, a *formula* or *string* or *tokening* of some members of a system (or "language") of elements for which there is a semantics or interpretation, and a provision (a mechanism of some sort) for reading or parsing the formula (1987:216 ; Dennett's italics).

By all the considerations I made above, I hope I made clear what could be a physically structured object that would explicitly encode some content: in order to catch the specificity of explicitly represented information in the brain, semantically distinct (physical) structures seem more realistic than (intrinsically and syntactically) structured physical objects. As I held, semantic discreteness is not necessarily mirrored by discrete and intrinsic syntax. Also, the importance of a user-understander of the content doesn't necessarily imply the explicit content should be able to be used or understood by the whole system it takes place within, and should ultimately be defined that way (against the procedural view).

Paying attention to the definition of explicit content should not lead us to think that explicit representations play a fundamental role in our cognitive systems. Explicitness of content, I hold, necessarily comes with implicitness and tacitness of content: by itself, a representational content cannot *do* something. Background knowledge, cognitive architectures, environmental constants, nested and exploited informations, computational transient processes are, amongst other things, the other representational parameters defining how explicitly represented contents are used in the system and how they can be the contentful objects of various dynamic cognitive attitudes[14].

Pierre Steiner

CEPERC, Université de Provence

[14] Various parts of this paper were presented in Paris (CREA, Research Group "CPER Représentationnalisme et Sciences Cognitives"), Aix en Provence (CEPERC) and Lund (ESPP Meeting 2005). Special thanks to Andre Abath for pressing me on various points.

Bibliography

BICKHARD, M.H. and TERVEEN, L. (1995), *Foundational Issues in Artificial Intelligence and Cognitive Science – Impasse and Solution*, Amsterdam, Elsevier Scientific.

CHURCHLAND, P.S. and SEJNOWSKI, T.J (1992), *The Computational Brain*, Cambridge (Mass.), MIT Press.

CHURCHLAND, P.M., and CHURCHLAND, P.S. (1998), *On the Contrary*, Cambridge (Mass.). MIT Press.

CHURCHLAND, P.S. (2002), *Brain-Wise. Studies in Neurophilosophy*, Cambridge (Mass.), MIT Press.

CLAPIN, H. (1999), "What, exactly, is explicitness?", *Behavioral and Brain Sciences*, 22:1, 150-151.

CLARK, A. (1992),"The Presence of a Symbol", *Connection Science*, Vol.4, Nos. 3 & 4, 115-129.

CLARK, A. (1993), *Associative Engines: Connectionism, Concepts and Representational Change*, Cambridge (Mass.), MIT Press.

CRICK, Fr., & Koch, Chr. (1995), "Are We Aware of Neural Activity in Primary Visual Cortex?", *Nature* 375, 121-123.

DENETT, D. (1987), *The Intentional Stance*, Cambridge (Mass.)/London, MIT Press.

DEVITT, M. (1991), "Why Fodor Can't Have It Both Ways", in Loewer and Rey (1991), 95-118.

DIENES, Z. and PERNER, J. (1999), "A Theory of Implicit and Explicit Knowledge", *Behavioral and Brain Sciences* 22:5, 735-808.

FODOR, J. (1982), *Representations*, Cambridge (Mass.), MIT Press.

FODOR, J. (1987), *Psychosemantics*, Cambridge (Mass.), MIT Press.

FODOR, J. (1987b), "A Situated Grandmother? Some Remarks on Proposals by Barwise and Perry", *Mind and Language*, vol.2, 1987.

FODOR, J. and PYLYSHYN, Z. (1988), "Connectionism and Cognitive Architecture: A Critical Analysis", *Cognition*, vol.28, 3-71.

FODOR, J. (1991), "Replies", in Loewer and Rey (1991), 255-319.

FODOR, J. (1994), *The Elm and the Expert*, Cambridge (Mass.), MIT Press.

KIRSH, D. (1990), "When is Information Explicitly Represented?", in Ph.Hanson (ed.), *Information, Language, and Cognition*, Vancouver Studies in Cognitive Science, vol.1, New York/Oxford, Oxford University Press, 340-365.

KIRSH, D. (2003), "Implicit and Explicit Representation", in N.Lynn (ed.), *Encyclopedia of Cognitive Science*, London, Nature Publishing Group, Vol.3, 478-481.

LOEWER, B. and REY, G. (ed.) (1991), *Meaning in Mind. Fodor and his Critics*, Oxford/Cambridge (Mass.), Blackwell.

LYCAN, W. (1986), "Tacit Belief", in R.Bogdan (ed.), *Belief: Form, Content, and Function*, Oxford, Clarendon Press, 61-82.

O'BRIEN, G. and OPIE, J. (1999), "A Connectionist Theory of Phenomenal Experience", *Behavioral and Brain Sciences*, 22:1, 127-196.

PYLYSHYN, Z. (1991), "Rules and Representations: Chomsky and Representational Realism", in A.Kasher (ed.), *The Chomskyan Turn*, Oxford/Cambridge (Mass.), Basil Blackwell, 231-251.

STICH, S. (1991), "Narrow Content Meets Fat Syntax", in Loewer and Rey (1991), pp.239-253.

VAN GELDER, T. (1990), "Why Distributed Representation is Inherently Non-Symbolic", in G.Dorffner (ed.), *Konnektionismus in Artificial Intelligence und Kognitionforschung*, Berlin, Springer Verlag, 1990, 58-66.