# Belief in Semantics and Psychology

by

Yen-fong Lau

B.A., physics and philosophy, Oxford University (1990)

Submitted to the Department of Linguistics and Philosophy
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Philosophy

at the

Massachusetts Institute of Technology

August 1994



Signature of Author . . . . . .
Department of Linguistics and Philosophy
22nd August 1994

Certified by . . . . . .
Professor Robert Stalnaker, Thesis Supervisor
Department of Linguistics and Philosophy

Accepted by. . . . . .
Professor George Boolos
Chairman, Departmental Committee on Graduate Studies

Belief in Semantics and Psychology

by

Yen-fong Lau

Submitted to the Department of Linguistics and Philosophy
on 22nd August 1994 in partial fulfillment of the requirements for
the Degree of Doctor of Philosophy in Philosophy

*Abstract*

This thesis consists of three papers discussing the nature of intentional mental states and their attribution, focusing on beliefs and thoughts. Chapter one looks at the problem of equivalence in giving a possible worlds semantics for belief reports. Proposed solutions by Paul Pietroski and Robert Stalnaker are examined and found to be unsatisfactory. I suggest that one might identify the denotation of a that-clause with an interpreted logical form, while holding a possible worlds theory of belief.

Chapter two discusses Martin Davies's *a priori* argument that thinking involves a language of thought. I point out that Davies's argument rests on an equivocation and is therefore invalid. Davies's argument appeals to considerations based on the possession of concepts. I explain why such considerations do not provide any *a priori* reason for accepting his conclusion.

In response to externalism, it has been argued that beliefs and thoughts have "narrow contents" that are determined by the intrinsic properties of their subjects. In chapter three I distinguish between three different explanatory tasks that narrow content is supposed to perform. I argue that none of the three motivations justify the thesis that all beliefs and thoughts have narrow content.

Thesis supervisor : Professor Robert Stalnaker

*Acknowledgement*

Many salient causes have jointly made this dissertation possible, and better than it would have been otherwise given the limited time I had. First and foremost I would like to thank my main supervisor, Bob Stalnaker. Bob has helped sort out many of my half baked ideas, and his ability to point out the deeper issues and what is at stake has provided a much needed sense of direction. As always, Ned Block has offered detailed and constructive comments that are extremely helpful. His advice, encouragement and good humour are very much appreciated. I have also learnt much from Professor Noam Chomsky. Noam's views have enabled me to see many philosophical issues in a new light, and made me rethink what philosophy is all about. I am most grateful to all three of them for various discussions. I have also learnt a lot from other members of the philosophy faculty. They each have their own special way of doing and teaching philosophy, which I can only hope to emulate. Discussions with fellow students over the years have been most enjoyable, and will certainly be missed. In particular, Josep Maciá has offered helpful comments on the first chapter, and Daniel Stoljar has provided comments on the third chapter that led to many improvements. On a less personal note, I would like to thank MIT and my department for financial support, though I hope the funding situation can improve further for future students. My deepest gratitude goes to my parents, for their love, care and support all through the years. This thesis is dedicated to them. I would also like to thank my marvellous friends Delman, David, Lam, Joseph and Sharifa for pulling me through some troubled times. Finally, thanks also to Lusina my love, for the life that we have shared, and what is to come.

# Belief in Semantics and Psychology

Table of Contents

# *Chapter One*

# Possible Worlds Semantics For Belief Sentences[1]

Whatever it is that a semantics for a natural language should explain, it should at least tell us what the truth-conditions of the sentences are. As David Lewis says, "semantics with no treatment of truth-conditions is not semantics." This paper is about possible worlds semantics for belief sentences in English. On the standard possible worlds approach, "believe" expresses a relation between things that have beliefs and propositions - abstract objects that are to be identified with sets of possible worlds. It has often been argued that this proposal suffers from what is called the problem of equivalence, and that propositions must be individuated more finely to provide the correct truth-conditions of belief sentences. The aim of this paper is to examine the attempts by Paul Pietroski and Robert Stalnaker to solve this problem. Although I shall conclude that neither are satisfactory, it is worth pointing out that giving up on the possible worlds theory of "believe" is quite compatible with holding a possible worlds theory of belief. At the end of the paper I shall explain what this position comes down to.

## 1. Recipe for disasters

The task at hand is to provide a systematic account of the truth-conditions of belief sentences of the form "*X* believes that *p*". Borrowing the linguist's terminology, I shall use "NP" for the subject term of the sentence, "CP" for the that-clause, and "IP" for

---

[1] Boldface in quotations in this and other chapters are added for emphasis and not in original text.

the embedded sentence. (Subscript indicates which particular sentence is the syntactic item to be taken from.) Let us now review the assumptions that lead to trouble for the possible worlds theorist. First, it is assumed that belief sentences should be given a binary analysis : a belief report[2] "*X* believes that *p*" is true if and only if the object denoted by the NP stands in the believe relation to the proposition denoted by the CP, an abstract object which is to be identified with a set of possible worlds.[3] Call this *the binary thesis.*

On the possible worlds theory, a proposition *p* is true at a possible world *w* if and only if *w* is a member of *p*. If *p* is true at the actual world, then it is true *simpliciter*. Logical relations among propositions can readily be defined in terms of set-theoretic relations among sets of worlds, e.g. a proposition *p* entails a proposition *q* if and only if *p* is a subset of *q*. The necessarily false proposition, i.e. the empty set ∅, thus entails every other proposition. But if propositions are sets of possible worlds, which one does a CP denote? According to what we might call *the intension thesis*, a CP denotes the intension of its embedded IP, i.e. the set of possible worlds with respect to which the IP is true.

Disaster struck immediately if a possible worlds theorist accepts the binary thesis and the intension thesis. For they imply that any two belief reports have the same truth-conditions, and so the same truth-value, if their NPs refer to the same individual and their IPs are necessarily equivalent (i.e. true and false with respect to the same set of possible worlds). This is what is known as *the problem of equivalence*. It is a problem because such a consequence is surely counterintuitive. For consider these two sentences :

---

[2] By a "belief report" I mean an assertive utterance (i.e. a token) of the corresponding belief sentence in some particular context. To avoid being cumbersome I shall be sloppy on the distinction between types and tokens though.

[3] One might also take propositions to be functions from possible worlds into truth values : a proposition *p* is true (or false) at a world *w* if $p(w)=T$ (or *F*). Unlike the set-of-worlds version, this proposal allows propositions which are neither true or false at a world. Nothing in the paper hangs on this feature so I shall adopt the simpler set-of-worlds version.

(1) Hilbert believes that 2=1

(2) Hilbert believes that arithmetic is complete

Both $IP_1$ and $IP_2$ are necessarily false, and so their CPs denote the empty set $\varnothing$. According to the current proposal then (1) and (2) have the same truth-value. But of course they might not. In this paper I shall assume that proper names are rigid designators, and thus (3) and (4) are also predicted to have the same truth value. Yet surely we can imagine a context where one but not the other is true :

(3) Lusina believes that Mark Twain wrote novels

(4) Lusina believes that Samuel Clemens wrote novels

Note that the problem of equivalence does not depend on how the possible worlds theorist might go on to analyse the believe relation. For the purpose of discussion, however, I shall adopt the following standard possible worlds theory of the believe relation. The proposal is that the propositions a subject $X$ believes are defined by a set of *doxastic possibilities*, $DOX(X)$, and $X$ believes a proposition $p$ if and only if $p$ is true at all the worlds in $DOX(X)$. In other words, $DOX(X) \subseteq p$. Of course, to assume this theory is not to say that this is unproblematic, since the theory implies that the propositions one believes are closed under entailment :

(5) If $X$ believes propositions $p_1 \ldots p_n$ and they entail $q$, then $X$ also believes $q$.

Call this the *strong closure principle*. On the face of it, the strong closure principle does not seem to be true of us at all. It has sometimes been proposed that we can avoid (5) by

postulating that what a subject believes are defined by more than one set of doxastic possibilities.[4] Thus one might believe $p$ because $DOX_1(X) \subseteq p$, and believes $q$ because $DOX_2(X) \subseteq q$, even though one does not believe their consequence $r$ because it does not follow that there is a $DOX_i(X)$ such that $DOX_i(X) \subseteq r$. Still, this alternative proposal entails the *weak closure principle*, an instance of (5) with $n$=1. Adding weak closure to the intention and the binary thesis is enough to generate *the problem of logical omniscience* : if the intension of "$p$" entails the intension of "$q$", and "$X$ believes that $p$" is true, then "$X$ believes that $q$" is also true. So for example if (6) is true, so is (7), which is surely wrong :

(6) Thales believed that the earth floats on water

(7) Thales believed that the earth floats on a compound that contains hydrogen

The problems of equivalence and logical omniscience show that our linguistic intuitions are incompatible with some of the principles accepted by the possible worlds theorist. So he either has to reject the validity of our linguistic intuitions, or he has to give up some of his principles of semantics and belief. Given the *magnitude* of the counterintuitive consequences, it is surely inadvisable to reject our considered intuitive judgements, which at this stage provide the best evidence we have for our semantic theory. Both problems arise, however, only if we accept (i) the binary thesis, that "believes" expresses a binary relation between the subject and what is denoted by the CP, and (ii) the intention thesis, that a CP denotes the intension of its embedded IP. If our linguistic intuitions are to be preserved, one of these two assumptions will have to go. Both Pietroski and Stalnaker accept (i) and reject (ii). Furthermore, both agree that a CP denotes a set of possible worlds, even though they disagree on *which* set it is. However, I

---

[4] See Chapter 5 of Stalnaker (1984) *Inquiry* Cambridge : MIT Press.

shall argue that their proposals are problematic, and that if (i) is to be kept, then they should identify the denotation of a CP with something *other* than a set of possible worlds.

## 2. Operation rescue I : The metalinguistic strategy

Let us begin by looking at Pietroski's proposal.[5] The leading idea is that in a *normal* context, the truth of a belief report "*X* believes that *p*" requires that *X* believes *both* the intension of "*p*" and some appropriate metalinguistic proposition. Pietroski does accept the strong closure principle, and so this is equivalent to requiring that *X* believes the *conjunction* of the intension and the metalinguistic proposition. He proposes that what a CP denotes is normally given by the following rule :

(8) *Pietroski's CP-rule* : a CP "that *p*" denotes the set of worlds where (a) it is true that *p*, and (b) some sentence similar to "*p*" is true.[6]

The relevant similarity relation is context-dependent, but is often something like the relation of *being a translation of*. Although whether something translates another is often context-dependent and vague, this is no criticism of the proposal because belief attributions do appear to be vague and context sensitive. To see how the proposal works, consider for example Pierre the monolingual Frenchman who believes that snow is white. His doxastic possibilities include only possible worlds where snow is white. Thus he believes the set { *w* : snow is white in *w* }. Pierre also sincerely assents to "La neige est blanche", so he also believes the proposition *S* which is the set { *w* : "La neige est blanche" is true in *w* }. By strong closure he believes :

---

[5] Paul Pietroski (1993) "Possible Worlds, Syntax, and Opacity" in *Analysis* October 1993, pp.270-280.

[6] Pietroski provides a formal recursive definition of the denotation of a CP. But since his proposal does not deal with demonstratives, the "disquotational" version (8) is what his proposal comes down to.

(9) { *w* : snow is white in *w* and "La neige est blanche" is true in *w* }

Since "La neige est blanche" translates as "snow is white", the two sentences are similar and thus (9) is a subset of (10) (and hence entails it) :

(10) { *w* : snow is white in *w* and something similar to "snow is white" is true in *w* }

Since Pierre believes (9), by weak closure Pierre believes (10) also. But according to Pietroski's CP-rule (10) is what is denoted by the CP of "Pierre believes that snow is white". So this belief report is indeed true in accordance with intuition.

Let's see how this proposal allows Pietroski to avoid the problem of equivalence. Consider Lusina the monolingual English speaker, for which (3) is true. So she believes the set :

(11) { *w* : Mark Twain (i.e. Samuel Clemens) wrote novels in *w* and something similar to "Mark Twain wrote novels" is true in *w* }

But it does not follow that (4) is true, since this requires her to believe :

(12) { *w* : Twain wrote novels in *w* and something similar to "Samuel Clemens wrote novels" is true in *w* }

But perhaps Lusina mistakenly thinks that "Samuel Clemens" refers to some New England patriot who brew beer but never wrote, so she sincerely dissents from "Samuel Clemens wrote novels" and any other sentence similar to it. There are no sentences similar to

"Samuel Clemens wrote novels" that she believes to be true. So she does not believe { *w* : something similar to "Samuel Clemens wrote novels" is true in *w* }. But by weak closure she would believe this set if she were to believe (12). So Lusina does not believe (12), and hence (4) is false. So (3) and (4) have different truth-values, even though their IPs are necessarily equivalent. Furthermore, note that Pietroski's CP-rule also avoids the problem of logical omniscience. Thus consider (6) and (7). Suppose that Aristotle were right in that Thales did believe that the earth floats on water. So (6) is true. But presumably there is no sentence in ancient Greek that translates as "the earth floats on a compound that contains hydrogen", and which Thales believed to be true. So Thales did not stand in the belief relation to { *w* : the earth floats on a compound containing hydrogen in *w* and something similar to "the earth floats on a compound containing hydrogen" is true in *w* }. So (6) and (7) can have different truth-values.

Note that the CP-rule (8) does not apply to all contexts. According to Pietroski, "abnormal" contexts (which need not be infrequent or unrealistic) in which the rule does not apply are of two sorts. First, there are contexts where we might drop (8b) from the revised CP-rule, thus in effect reverting to the original intension thesis, where the CP denotes the intension of the IP. One reason for this move is to allow attributing beliefs to organisms that do not have a language. Fido the dog might believe that there is a bone in the yard, but surely there are no sentences which it believes to be true, and which is similar to "there is a bone in the yard". So in such an "abnormal" context, if we drop (8b), the CP of "Fido believes that there is a bone in the yard" now denotes the intension of the embedded sentence, the set of worlds in which a bone is in the yard. The attribution will be true if Fido stands in the belief relation to this set, even if Fido bears no attitude relation to any sentences whatsoever.

The second type of abnormal contexts are those where (8a) is dropped instead. So in such contexts it is sufficient for "*X* believes that *p*" to be true if there is some sentence similar to "*p*" that *X* believes to be true. One reason for this move is to deal with cases like (1) and (2). $IP_1$ and $IP_2$ are both necessarily false and so they express the same proposition ∅. But on a normal context reading according to (8) the CPs denote the proposition that is the conjunction of ∅ and some other metalinguistic proposition. But the conjunction of ∅ with any proposition is of course ∅ itself. So (1) and (2) always have the same truth-value, which is obviously false. Pietroski's way out is to say that the attribution of mathematical beliefs are abnormal cases where the proposition expressed by the IP "is of no interest". In all such cases we "ignore" and drop requirement (8a). So $CP_1$ and $CP_2$ denote respectively,

(13) { *w* : something similar to "2=1" is true in *w* }

(14) { *w* : something similar to "arithmetic is complete" is true in *w* }[7]

Suppose Hilbert is a monolingual German speaker who dissents from "zwei ist eins" and other sentences similar to "2=1". He therefore does not believe (13). But Hilbert does think that all theorems of arithmetic can be proved. There is a German sentence *s* that he believes to be true, which translates into English as "arithmetic is complete". So Hilbert believes { *w* : *s* is true in *w* }, which is a subset of (14). So Hilbert believes (14) also by weak closure.Thus (2) is false while (3) is true, consistent with our intuition that they have different truth-conditions.

---

[7] Note that strictly speaking "similar" in (12a) and (12b) should be understood as "actually similar". For presumably the similarity relation (like its instance *being a translation of*) preserves truth value, and so there cannot be a world with a true sentence *s* such that *s* in that world is similar to "arithmetic is complete" in the actual world. More precisely then, the metalinguistic proposition denoted by "that *p*" in such contexts should be { *w* : there is some sentence *s* that is true in *w* and which in the actual world is similar to "*p*" }. I hope there would be no confusion by sticking to the simpler formulation.

## 3. Problems and Counterexamples

To sum up, Pietroski's proposal is that normally "that p" denotes the conjunction of the intension of "p" and a metalinguistic proposition. But there are abnormal contexts where one or the other component proposition is left out. Pietroski thinks that his three-tier proposal is readily explicable "on the plausible assumption that we typically ascribe beliefs to explain behaviour" :

> For we do not have to explain *assent* when it comes to dogs; and response to linguistic entities is the only behaviour that could prompt ascriptions [of mathematical beliefs] ... But we normally ascribe beliefs to language users, whose 'assenting behaviour' is usually part of that which we explain by such ascription. That is, we normally take metalinguistic and nonlinguistic beliefs into account when saying how *speakers* 'take the world to be.'[8]

Should the possible worlds theorist rests with such an account? I think not. First, despite his justification for his three-tier semantics for interpreting a CP, he has yet to explain what is it that determines which of the three semantic rules to use in a particular context of belief attribution. We know that an abnormal context is one where either (8a) or (8b) does not apply, but what is it about a context that makes it abnormal? Pietroski does suggest that contexts where (8a) does not operate include those where mathematical beliefs are attributed. But are there other cases? Presumably, this would include any instance of "*X* believes that *p*" where the IP is necessarily false (or else these reports will all be false on a normal-context reading). So should we say that abnormal contexts where (8a) is set aside are ones where the IP is necessary? But then what about "the confused undergraduate student believes that there are no thinking beings"? The embedded sentence

---

[8] Pietroski *op. cit.* page 278.

is arguably a contingent one. But presumably, any world in which nothing thinks are ones where no languages are spoken. But then on a normal context reading the CP of the sentence will denote the empty set, since in any mindless world there won't be anything that is true and which is actually similar to "there are no thinking beings". Perhaps Pietroski would say that this is also a case of abnormal context, so that the CP denotes a contingent proposition. But why should it? Sometimes Pietroski speaks of "ignoring" (8a), or "taking into account" (8b). Does this mean that whether a context is normal or not is a matter of the speaker's intention? But this cannot be right. For surely I cannot make (1) and (2) come out to have the same truth-condition, simply by intending that the context is normal and that (8a) applies. But if the normality of a context (or the lack of it) is not determined by the speaker's intention, then what else is relevant? Is it a matter of tacit conventions? But whether this is correct depends on what the content of those conventions might be. We still need to fill in the blank in "it is a matter of convention that a context with feature _____ is a normal context". What Pietroski has to provide are motivated principles that distinguish normal from abnormal contexts, other than that "normal" is when the revised CP-rule works and "abnormal" is when it doesn't.

The need for some principled distinction between normal and abnormal contexts would be less pressing if belief reports can always be assigned the correct truth-condition by one of the three semantic rules. Unfortunately this is not the case. Consider for example our monolingual German mathematician Hilbert, for which "Hilbert believes that arithmetic is complete" is true. But now suppose Hilbert is told that "arithmetic is incomplete" is a true sentence in English, but he has no idea what it means. Hilbert still believes that arithmetic is complete, but taking his informer to be trustworthy, he comes to acquire the new and true belief that "arithmetic is incomplete" is a true English sentence. So he now believes the set $K$, which is { $w$ : "arithmetic is incomplete" is true in $w$ }. Consider "Hilbert believes that arithmetic is incomplete", which of course is false.

However, all three semantic rules for interpreting a CP predict that it is true. First, if we suppose that this is an abnormal context where (8b) does not apply, then the CP denotes the necessarily true proposition. Since Hilbert believes $K$ and $K$ entails the necessarily true proposition, by weak closure he believes the latter also. On the other hand, taking either the normal context reading where both (8a) and (8b) apply, or the abnormal context reading that drops (8a), the CP denotes the same metalinguistic proposition $L$, which is { $w$ : something similar to "arithmetic is incomplete" is true in $w$ }. Regardless of what the contextually relevant similarity relation is in the context, surely it has to be reflexive and so $K$ is a subset of $L$. Hilbert believes $K$ and by weak closure again he also believes $L$. Thus "Hilbert believes that arithmetic is incomplete" is true. Not only that, presumably the following is also true because both conjuncts are true : "Hilbert believes that arithmetic is incomplete and he believes that arithmetic is complete". A truly disastrous consequence indeed.

It will not do to respond by stipulating that similarity has to be non-reflexive. For presumably there are other non-German (eg. French) sentences which are distinct from but are similar to "arithmetic is incomplete". If Hilbert believes any of them to be true the same result follows. It will also not do to respond by rejecting the weak closure principle, since it plays an essential role in Pietroski's proposal in attributing beliefs to non-English speakers (Recall the Pierre example on page 3). Note also that this objection is not restricted to attributions of mathematical beliefs. Similar counterexamples can be constructed with belief reports whose IPs are necessary. For example, intuitively, "Kathrin believes that diamonds are made of carbon" might be false, even if Kathrin is a monolingual German speaker who correctly believes that "diamonds are made of carbon" is (necessarily) true. In fact, more complicated counterexamples involving attributions of contingent beliefs can be constructed as well. Thus consider Lusina the monolingual English speaker again who believes that Twain wrote novels but that Clemens did not.

Suppose she comes to believe of some sentence in French that it is true, but which unbeknownst to her translates into English as "Clemens wrote novels". (Perhaps she mistakenly thinks it means Clemens wrote *no* novels.) So she believes both { *w* : Clemens wrote novels in *w* } and { *w* : something similar to "Clemens wrote novels" is true in *w* }. On Pietroski's proposal, "Lusina believes that Clemens wrote novels" is now true, even though intuitively it still is false.

It is of course quite obvious what the problem is. An underlying assumption behind the proposal is that necessarily equivalent beliefs are to be distinguished by their being dispositions to assent to different sentences. But assent is however not a good model for belief, for the simple reason that one can assent to a sentence without understanding it. Thus one common feature of all these counterexamples is that the subject comes to believe the proposition denoted by the CP by believing, of some sentence they do not understand, that it is true. Having pointed this out, one might think that the simple response is that assent without understanding should not be sufficient for believing the metalinguistic proposition. So perhaps Pietroski might modify the CP-rule slightly, say :

(15) The CP of "*X* believes that *p*" normally denotes the set worlds where (a) it is true that *p*, and (b) some sentence that *X* actually understands, and which is similar to "*p*", is true.

This proposal does seem to block the Hilbert counterexample. For "Hilbert believes that arithmetic is incomplete" to be true, Hilbert has to believe { *w* : some sentence *s* which Hilbert actually understands and is similar to "arithmetic is incomplete" is true in *w* }. This condition obtains only if there is a sentence similar to "arithmetic is incomplete" and that Hilbert understands and believes to be true. But there is no such sentence. Hilbert does believe that "arithmetic is incomplete" is true, but he does not understand it. The

German translation of this sentence is presumably one that he understands, but he does not believe that it is true. The same goes for the other counterexamples.

However, the new proposal does have *a lot* of objectionable consequences, of which I shall mention two. First, what is it to understand a sentence *s* is none too clear. But if it requires knowing the meaning of the words that occur in the sentence, then it would seem that (15) is too strong. For we do in fact attribute beliefs to subjects even when they have a mistaken belief as to the meaning of a word in the embedded sentence. Familiar examples include "Francine believes that a fortnight is a period of ten days", "Francine believes that Bill has arthritis in his thigh". In both cases, we might say that the English speaker Francine does not understand the embedded sentence of the report because there is some word in the sentence whose meaning she does not know. Yet both belief reports can still be true. To take such cases into account, we might modify (15) slightly by replacing "understands" with "partially understands". So "Francine believes that a fortnight is a period of ten days" is true if Francine partially understands the IP and believes it to be true. But now this opens the floodgate for cases where a subject *X* partially understands "*p*", but where we are reluctant to accept as true "*X* believes that *p*". Returning to the Hilbert case again, would Hilbert count as partially understanding "arithmetic is incomplete", if he understands every word in it except that he thinks "incomplete" means *complete*? But surely this is still not sufficient for "Hilbert believes that arithmetic is incomplete" to be true. It is hard to see what notion of understanding we might use to distinguish between those cases where misunderstanding is nonetheless sufficient for the belief (the Francine case) from those where misunderstanding does not (the Hilbert case).

But supposing that there is such an account, still (15) or any such modification will not offer the right semantics for a CP. For whatever notion of partial understanding we might employ, it will be the case that what a CP denotes in a belief report will vary

dramatically depending on who the subject is. If Jason is a monolingual English speaker, then the CP of "Jason believes that arithmetic is incomplete" will denote a set of worlds where there is a true *English* sentence actually similar to the IP and understood by Jason (fully or partially in whatever appropriate way. I shall ignore the qualification henceforth). On the other hand if Kathrin is a monolingual German speaker, then the same CP in "Kathrin believes that arithmetic is incomplete" will denote a *different* metalinguistic proposition, the set of worlds where there is a true *German* sentence that is actually similar to the IP and understood by Kathrin. It does not seem to be an intuitive proposal at all, since one would have thought that it is *the very same thing* that Jason and Kathrin are said to believe in these reports. Furthermore, what about belief reports with quantified NPs? Take for example (16) :

(16) Every logician believes that first order logic is complete

What does $CP_{16}$ denote? It cannot be the set of worlds where there is a true sentence that in the actual world is similar to the IP and understood by every logician. This is because the logicians who all believe that first order logic is complete might include monolingual speakers of different languages. There might be no single sentence that all of them understand and believe to be true, and which is similar to the IP. In which case $CP_{16}$ will denote the empty set, and so this predicts that (16) is false when it might not be. The only alternative in line with the current proposal is that the CP does not denote any proposition at all, that it conceals a bound variable. Say (16) should be analysed as :

(17) $\forall x$ ( $x$ is a logician $\rightarrow$ $x$ believes { $w$ : there is a true sentence in $w$ that is actually similar to "first order logic is complete" and which is understood by $x$ } )

So $CP_{16}$ has no denotation any more than "his mother" denotes in "every man loves his mother". But this just seems to be the wrong analysis. It is for example valid to infer from (16) to "There is something that every logician believes", which surely is to be analysed as :

(18) $\exists p \; \forall x \; ( x \text{ is a logician} \rightarrow x \text{ believes } p \; )$

(18) however, does not follow from (17), just as "there is something that every man loves" does not follow from "every man loves his mother".[9] I think this is good reason for thinking that the bound variable reading of (16) cannot be right. It does not seem to me that there are other alternatives for dealing with this problem, and which is in line with our modified version of Pietroski's proposal. So let us turn to Stalnaker's proposal instead.

## 4. Operation rescue II : the diagonalization strategy

In a series of publications, Robert Stalnaker has defended the viability of a possible worlds semantics for belief attributions.[10] Like Pietroski, Stalnaker accepts the binary analysis of belief sentences : [11]

---

[9] One might argue that (18) does follow from (17) because everyone believes the necessarily true proposition and so if (17) is true then (18) has to be true also. But we might replace "every" with "all and only" in (16) and (18) and the problem still comes up.

[10] See for example Stalnaker (1987) "Semantics for Belief" in *Philosophical Topics* Volume XV, No. 1, pp.177-190, and Stalnaker (1988) "Belief Attribution and Context" in Grimm and Merrill (eds.) *Contents of Thoughts* Tucson : University of Arizona Press.

[11] There is one difference though : Stalnaker thinks that what the CP denotes is the *proposition expressed* by the IP, which need not be its intension. Pietroski, on the other hand, thinks that the proposition expressed by the IP is just the intension, but that what the CP denotes is normally not the proposition expressed. Now one might wonder whether there is a real dispute here given that in any case both agree that the CP does not always denote its intension. This of course depends on what the notion of *the proposition expressed* is supposed to explicate. Sometimes the proposition expressed by an assertive utterance is supposed to be the information that is conveyed by the utterance. Sometimes

> the transitive verb *believe* expresses a relation between a person or other animate thing denoted by the subject term and a proposition denoted by the sentential complement that is the object of the verb. "Phoebe believes that fleas have wings" seems to say that Phoebe stands in the belief relation to the proposition *that fleas have wings*. I think that the semantics of belief really is as simple as it seems.[12]

What is complicated however, is the way in which the denotation of the CP is dependent on the background presuppositions against which the belief reports are made. The proposition that a CP denotes is the proposition expressed by its embedded IP, but *which* proposition it expresses is heavily context-dependent. To solve the problem of equivalence one has to understand the interaction of assertive utterances with the presuppositions made in the course of a conversation. According to Stalnaker, the mutual presuppositions that participants make define *the context set* - the set of possible worlds compatible with what are mutually presupposed. The set defines the range of possible ways the world can be that are "left open" by what the participants of the conversation take themselves to know. What an informative assertion does then is to *reduce* the context set, to eliminate further those possibilities that according to the speaker are not the way the world is. The principles for interpreting an assertive utterance are therefore guided by the assumption that the speaker intends to succeed in this task. As Stalnaker puts it,

> in general, to express a proposition is to select a subset of possible situations given by the context. This will be true of embedded sentences as well as sentences uttered on their own. For embedded sentences, we need an embedded, or as I will call it, a *derived context.*[13]

---

it is taken to be what the speaker says in making that utterance, or perhaps what the speaker represents himself as believing. But on all these (rough) accounts it would seem that the intension of the utterance is not a good candidate to be identified with the proposition expressed.

[12] Stalnaker, "Belief Attribution and Context", page 140.

The derived context (relative to the subject to which belief is attributed) is what is presupposed to be the set of doxastic possibilities for the subject. In other words, it is the set of possible worlds presupposed to be compatible with what the subject believes. This defines the information that participants of a conversation have as to what the subject believes. A belief report is an informative one then, if it succeeds in providing further information about the subject's doxastic possibilities, by identifying possible worlds in the derived context that the speaker takes to be incompatible with what the subject believes. Accepting an informative belief report as true would then have the consequence of reducing the derived context by eliminating such worlds. This is the framework that Stalnaker employs to deal with the problem of equivalence. The idea is that in interpreting an utterance we appeal to principles like an assertive utterance has to be informative, and so the proposition expressed should succeed in reducing the context set. Such principles are supposed to explain how utterances can express different propositions even though they have the same intension.

To see how the explanation goes, let us look at an example adapted from one of Stalnaker's papers. Consider a context in which Stalnaker and Daniels are talking about a certain person O'Leary. As it happens, O'Leary is someone who knows little about astronomy, and he mistakenly thinks that the planet that appears in the evening sky, which he knows people call "Hersperus", refers to the same planet as "Mars". Assume for the sake of simplicity that there are only two such worlds and call them *b* and *c*. Let *a* be the actual world where "Hersperus" refers to Venus and "Mars" to Mars. Now suppose that in the conversation both *a*, *b* and *c* are presupposed to be compatible with what O'Leary believes, i.e. the derived context is { *a*, *b*, *c* }. Daniels knows what O'Leary believes and to convey his knowledge to Stalnaker he says :

---

[13] Stalnaker, *op. cit.*, page 146.

(19) O'Leary believes that Hersperus is identical to Mars

Given the information we have about O'Leary, (19) is true. But the intension of its IP is of course the empty set, which nobody believes. So presumably the IP cannot denote its intension. Let us see how diagonalization might give the correct truth-condition to the belief report. Consider that particular token *t* of the IP that Daniels produced in making his belief report. We know that *t* has ∅ as its intension in the actual world. But we can also ask what intension *t* would have, *if it were to exist at other possible worlds*. This would allow us to define a certain function from possible worlds into propositions, which Stalnaker calls a *propositional concept*. More specifically, it is the function that maps a world *w* to the intension that *t* has at *w*. We might then represent the propositional concept with the following table :

| | *a* | *b* | *c* |
|---|---|---|---|
| *a* | *F* | *F* | *F* |
| *b* | *T* | *T* | *T* |
| *c* | *T* | *T* | *T* |

The leftmost column represents the possible worlds in the derived context. For each world *w* on the lefthand column, the row of truth values to its right represents the intension of *t* at *w* by showing whether the intension is true or false at a certain world. The intension of *t* is ∅ at the actual world *a*, and so it is false at all three worlds in the derived context as indicated by the top row of *F*s. However, if *t* were to be exist at worlds *b* and *c* it would have a different intension. The reason is that those are the possible worlds compatible with what O'Leary believes, where "Hersperus" and "Mars" corefer. At those worlds the intension of *t* is the necessarily true proposition, hence the two rows of *T*s next to *b* and *c*.

Given such a propositional concept, what can we say about the proposition that O'Leary is supposed to believe according to Daniels? This is the point where we appeal to the earlier suggestion that to express a propsition is to reduce the relevant set of possibilities in some appropriate way. If Daniels's belief report is to be informative, then it can't be that the IP expresses the necessarily false proposition, because that is to say that none of the possible worlds in the derived context is compatible with what O'Leary believes, which of course is not what Daniels is saying. On the other hand, the IP cannot be expressing the necessarily true proposition that is its intension at worlds *b* and *c*. So none of the *horizontal* proposition at each row of the table can be identified with what the IP expresses.

What Stalnaker proposes is that the proposition expressed by the IP is not any of the horizontal proposition. Instead it is identical to the what he calls the *diagonal proposition* of the propositional concept. Given a propositional concept *P*, the diagonal proposition is simply the set of worlds { *w* : *P* maps *w* to a proposition true at *w* }. The diagonal proposition is so named because it can be "read off" from the diagonal of the table. It includes exactly those worlds where the horizontal proposition is true at that world. In this particular example, the diagonal proposition is thus the set { *b* , *c* }. According to Stalnaker, this is the proposition that is expressed by the IP of Daniels's utterance. Since { *b* , *c* } is true at O'Leary's doxastic possibilities, Daniels's belief report is therefore true, and this is indeed correct. Furthermore, on Stalnaker's theory, to accept a belief report is to accept that the possible worlds incompatible with what the IP expresses are not among the subject's doxastic possibilities. So we expect that if Stalnaker accepts Daniels's report, the derived context will change from { *a*, *b*, *c* } to { *b*, *c* }, and again this seems right.

On the other hand, suppose that Daniels were to utter (20) instead of (19) :

(20) O'Leary believes that Hersperus is not identical to itself

At each possible world in the derived context, $IP_{20}$ will have the same intension, namely the empty set. The propositional concept defined for the IP is therefore one that maps every world in the derived context to ∅. The diagonal proposition is thus also the empty set, which O'Leary does not believe. Thus (19) is true while (20) is false. So the diagonalization procedure does succeed in giving different truth-conditions to the two belief reports, even though their IPs have the same intension in the actual world.

## 5. Problems with Diagonalization

I now want to consider how successful diagonalization is as a general strategy for dealing with the problem of equivalence. It has to be pointed out that Stalnaker never claims that the proposition expressed by an IP is *always* given by its diagonal proposition. This of course raises the question of when diagonalization is supposed to apply, and when it does not. Without an answer it will be difficult to evaluate the account. What I shall do is to focus on cases where it is not plausible to take the proposition expressed by the IP to be its intension, and see if diagonalization might offer a better account. We will see that diagonalization faces serious problems that are quite similar to those faced by Pietroski's proposal.

Let us focus on how the diagonal proposition expressed by the IP is defined. Since the value of the diagonal proposition is a function of the propositional concept, the crucial question is how the propositional concept is to be defined. As suggested earlier, Stalnaker's proposal is that the propositional concept is defined by applying a counterfactual procedure, as illustrated in this example of his :[14]

[14] Stalnaker, "Semantics for Belief", page 187.

> We ask something like the following question : **If Daniels were to utter the sounds he is uttering in a possible world compatible with O'Leary's beliefs, what would the content [i.e. intension] of those sounds be**? If the solar system were arranged so that Mars appears in the evening where Venus in fact does, then Daniels and I, as well as O'Leary, would use the name *Hersperus* to refer to Mars. And so, according to the semantical rules in that world, Daniels' sentential complement *that Hersperus is Mars* expresses a necessary truth. If we extend the propositional concept in this way, defining it for the situations that might, for all Daniels and I are presupposing, be compatible with O'Leary's beliefs, then the diagonal of that propositional concept will be the proposition that seems, intuitively, to be one O'Leary is said to believe.

So the general idea seems to be this : the value of the propositional concept at a possible world *w* is defined by the intension that an utterance of the IP will have at *w* according to the semantic rules for the IP in that possible world.

It seems to me that such a proposal faces serious problems in accounting for belief reports where the subjects do not speak English. The reason is this : suppose we have a subject who does not speak English and in particular *he has no beliefs as to what the correct semantic rules for the IP might be*. There is then a wide range of possible semantic rules for the IP, any of which could have been the correct one for all that the subject believes. We might suppose that one of them, call it *R*, determines that the IP has the necessarily false proposition as its intension. There surely is no reason why there cannot be a subject whose beliefs are compatible with this possibility. But if this is right, it then follows that there is a doxastic possibility of this subject where *R* is the correct semantic rule for the IP at that world. If the value of the propositional concept at such a world is given by the intension of the IP as determined by the semantic rules there, then

we will have to conclude that the diagonal proposition is false at *w*, and so false at one of the subject's doxastic possibilities. But recall that on the standard possible worlds account, an individual believes a proposition *p* if and only if *p* is true at *all* of that individual's doxastic possibilities. So this counterfactual procedure predicts that the belief report is false solely on the basis that the subject does not speak English! Surely this cannot be right.

In discussing the O'Leary example, Stalnaker does consider the possibility that O'Leary might not speak English, but he thinks that it does not create any special problem for diagonalization :

> What if O'Leary speaks some language other than English? That will make no difference to the explanation, so long as he has some acquaintance with Venus as it appears in the evening, either through having seen it, or through having acquired some name that denotes Venus because Venus appears where it does in the evening. The propositional concept we construct is the one not for the sentence as O'Leary would use or understand it, **but for the sentence as the speaker and addressee would use and understand it if they were in the possible worlds relative to which the propositional concept is being defined.**[15]

It is indeed true that the counterfactual procedure does not depend on the subject being able to speak English. But the problem I have raised arises precisely because of the *lack* of beliefs about English on the part of the subject, who has no opinion on how the speaker and addressee use and understand the IP. Note that the subject does not need to have any bizzare beliefs about English at all. The point is, given his lack of beliefs as to the correct usage of the IP, there might well be possible worlds compatible with what he believes, where the speaker and the addressee use and understand the IP in all sorts of ways that

---

[15] Stalnaker, *op. cit.*, page 187.

deviate from normal English usage. At such doxastic possibilities, the IP will be necessarily false. So for example, consider a possible world *u* where Daniels and Stalnaker use and understand "Hersperus is Mars" as meaning 2=1 for example. Since O'Leary has no opinion on what "Hersperus is Mars" means, this possibility is indeed compatible with what he believes. So *u* is one of O'Leary's doxastic possibilities. If as Stalnaker suggests we ask how the sentence *would be used and understood* by the speaker and addressee if they *were* at such a world, then surely this is none other than how they *in fact* use and understand it there, namely as a sentence that means 2=1. We are still forced to the conclusion that the belief report is false when it need not be.

One might say that the intensions of the IP as used and understood in such deviant ways are not appropriate for defining the propositional concept. Perhaps the value should be defined by the intension that the IP will have in *w*, if the way it is used and understood in *w* is *the same as the way it is used and understood in the actual world*. But what count as sameness of use? Daniels and Stalnaker do actually use "Hersperus" to refer to Venus, and "Mars" to Mars, and suppose they both understand that the sentence is necessarily false. If they were to use and understand it in *w* the same way they actually do, i.e. with "Hersperus" referring to Venus, etc, one would think that the IP will still have ∅ as its intension in *w*. Again the belief report (19) is predicted to be false when it might not be.

We might perhaps sum up the problem this way. To carry out diagonalization, we need to know how to define the value of the propositional concept at a doxastically possible world *w*. The current proposal has it that the value is identical to the intension that an *appropriate* token of the IP has at *w* if it were to exist at that world. However, the problem is that we have no account of what appropriateness comes down to. As we have seen, being a token of the IP is not sufficient. Nor is appropriateness a matter of being

used and understood the same way as in the actual world. But without any further account, we would have little reason to think that diagonalization can provide the correct truth-conditions of belief reports.

## 6. Modes of Presentation

But all is not lost. In the earlier passage where he considers the possibility that O'Leary does not speak English, Stalnaker claims that this does not affect diagonalization "so long as he has some acquaintance with Venus as it appears in the evening, either through having seen it, or through having acquired some name that denotes Venus because Venus appears where it does in the evening."[16] This remark seems to suggest that acquaintance somehow enters into diagonalization. But how?

As was originally introduced by Russell, being acquainted with some object is supposed to be necessary for having thoughts about that particular object, or being able to refer to it. I might sincerely assert that the richest person on earth owns at least a dozen luxurious mansions. But if I am not acquainted with whoever it is that is the richest person on earth, I could not be referring to or having a thought about any particular individual. It is of course not very clear what it is to be acquainted with something. Does it involve some kind of causal relation, in virtue of which certain mental states of the subject are causally dependent on that which he is acquainted with, or does it require possession of "discriminating knowledge", knowledge which enables the subject to distinguish the object in question from other ones?[17]

---

[16] Stalnaker, *op. cit.*, page 187.

[17] In "Belief Attribution and Context", Stalnaker argues that a weak causal relation can suffice for acquaintance. The idea that discriminating knowledge is necessary is explained and defended in Gareth Evans (1982) *The Varieties of Reference* Oxford : Oxford University Press.

Whatever the details of the account are going to be, I shall assume that if a subject *S* is acquainted with some object *o*, there is then some property of *o* which explains why *S* is acquainted with *it* and not with some other object. (So this property might be the property of *being the unique individual that is causally related to the subject in such and such a way*, or *being the unique individual that a certain body of discriminating knowledge K is true of*.) For obvious reasons, I shall call such a property the *mode of presentation* (MOP for short) that *S* uses to pick out *o*. I leave open the possibility that *S* uses more than one MOP to pick out the same object.

The general idea is that a MOP specifies *a way in which* the subject is acquainted with some object. But how might this help the diagonalization procedure? Here is a proposal. Suppose that associated with the use of a referring term is a MOP that the speaker uses to pick out the referent of the term. Now consider an assertive utterance of "*X* believes that *p*" by a speaker *S*. On this proposal, associated with the utterance of "*p*" are MOPs that pick out the actual referents of the constitituent referring expressions. To define the value of the propositional concept of "*p*" at *w*, we carry out the following counterfactual procedure. First we ask, if an utterance of "*p*", associated with the same MOPs, is made by the speaker in *w*, what would those MOPs pick out at *w*? Take the objects thus picked out to be the referents of the constituent referring expressions of "*p*" in *w*. This determines the intension of that specific utterance in *w*. Proposal : this is the value of the propositional concept at *w*.

The basic idea behind this proposal is that the value of the propositional concept is not defined by the intension of any arbitrary utterance. Instead it is given by the intension of an utterance where the speaker is acquainted with the referents the same way as in the actual world. In the O'Leary example, both O'Leary and Daniels are acquainted with Venus as it appears in the evening, but O'Leary thinks that the planet so acquainted is

Mars. At O'Leary's doxastic possibilities, the MOP associated with Daniels' utterance of "Hersperus" will presumably pick out Mars, even at worlds where "Hersperus is Mars" means 2=1. But if O'Leary is not acquainted with Venus as it appears in the evening, then it is not clear what the MOP will pick out at O'Leary's doxastic possibilities. Presumably in such a case (19) will be false. So perhaps this explains Stalnaker's remark that the diagonalization explanation is not affected by whether O'Leary speaks English at all, as long as he is acquainted with Venus as it appears in the evening.

But it seems to me that this remark also brings out a shortcoming of this approach : for a speaker $S$ to succeed in attributing a belief about $o$ to a subject $X$, it has to be the case that both $S$ and $X$ are acquainted with $o$ in the same way. But this seems to make successful belief attribution much more difficult and speculative than it really is. Consider for example Jason's belief report "Kathrin believes that diamonds are made of carbon". Since diamonds are indeed carbon crystals in a special lattice structure, the IP of this belief report is necessarily true. So I shall assume that the proposition expressed by the IP is not its intension, or else this would make the belief report true regardless of what Kathrin believes. How might diagonalization give us the correct truth-condition of the report? On the current proposal, to define first of all the propositional concept of the IP, one would have to appeal to the MOP that Jason actually uses to pick out diamonds, and consider what it picks out at Kathrin's doxastic possibilities. But what if Kathrin is a monolingual German speaker who calls diamonds "diamant", and is acquainted with diamonds very differently from Jason, a monolingual English speaker? Perhaps Jason thinks of diamonds as expensive gemstones sold in jewellery shops made of the same kind of material as pencil leads, whereas Kathrin's information about diamonds derive exclusively from some limited scientific sources that do not mention the ornamental function of diamonds. Perhaps she has never even encountered such things as pencils, and maybe she has no opinion on what is sold in jewellery shops either. But surely this does

not prevent her from having beliefs about diamonds. She might know that diamonds are carbon crystals in some kind of tetrahedral lattice structure, and that it is the hardest substance on earth. Given Kathrin's limited information, it is thus compatible with what Kathrin believes that jewellery shops do not sell stuff called "diamant", but they do have some carbon-free gems called "diamonds" in English, and which contain stuff that are also used in making writing instruments. One would think that at those of Kathrin's doxastic possibilities where this is in fact the case, the MOP that Jason actually uses to pick out diamonds will pick out the carbon-free gems instead of what Kathrin calls "diamants". If the intension of the IP "diamonds are made of carbon" at these possible worlds are defined in terms of the referents picked out by Jason's MOPs, then obviously we would expect the resulting diagonal proposition to be false at some such doxastic possibilities. The belief report will turn out to be false even though intuitively it is not.

There is however an obvious response available. The value of the propositional concept at $w$ is supposed to be defined by the intension of an appropriate IP at $w$. If the intension is not defined in terms of the MOPs that the *speaker* associates with the IP, perhaps we might appeal to the MOPs used by the *subject*? Here is a proposal : given an IP, consider the actual referents of the constituent referring terms and ask which are the MOPs that the *subject* use to pick them out. These then are the MOPs which determine the referents of those terms at the subject's doxastic possibilities. So take whichever MOP Kathrin actually uses to pick out diamonds and carbon. Call them respectively $D$ and $C$, and suppose that $d_w$ and $c_w$ are the substances they pick out at a doxastic possibility $w$. We can then ask the further question of what intension the IP has at $w$ if $d_w$ is the referent of "diamonds", $c_w$ is the referent of "carbon", etc.. The answer will be the value of the propositional concept at $w$. If Kathrin does believe that diamonds are made of carbon, we will expect that at all of her doxastic possibilities, $D$ picks out something called "diamant", and across all such worlds they are made of stuff picked out by $C$. The

diagonal proposition will then be true at all of her doxastic possibilities. Jason's belief report will be true even though he and Kathrin are acquainted with diamonds in very different ways.

However, what if the subject has more than one MOP which picks out the same object? Suppose that O'Leary is acquainted with Venus both as it appears in the evening and as it appears at dawn, but he does not know that it is one and the same planet. So suppose that (21) is true but (22) is false :

(21) O'Leary believes that Hersperus is Mars

(22) O'Leary believes that Phosphorus is Mars

On the present proposal, the value of the propositional concept of the IP at a world *w* is defined by the intension of an appropriate token of the IP at *w*, and the appropriate token is one associated with the MOPs that the subject actually use to pick out the actual referents of the IP. If (21) is true but (22) is false, the propositional concepts for the two IPs will have to differ in value at some doxastic possibility *w*. That is, the appropriate tokens of the two IPs will have different intensions at *w*. This implies that $IP_{21}$ and $IP_{22}$ must be associated with different MOPs that pick out different referents at *w*. But the actual referents of the two IPs are the same, so what is it that explains the different associations of MOPs? It cannot be arbitrary since presumably it is the different associations that explain why (21) is true but (22) is false, and not the other way round.

Perhaps there is some procedure in terms of contextual salience that selects, among those MOPs used by a subject to pick out the same thing, which is the right one to be associated with the IP in defining the propositional concept. I shall not be concerned with the details of such an account here. Not that it is straightforward, but because I think

there is a more serious problem facing the current proposal, regardless of how this account of contextual salience might be developed. The problem is one that afflicts Pietroski's proposal as well. On both accounts, what a CP denotes is much too subject-dependent to account for people believing the same thing and certain patterns of inference relating the objects of beliefs. In the next two final sections I shall explain how this problem arises for Stalnaker's proposal and suggest a possible response.

## 7. Accounting for Validity

In discussing Stalnaker's proposal, I argued first of all that diagonalization using the simple counterfactual procedure gives the wrong truth-conditions. One way out of the problem is to introduce modes of presentations. Their job is to pick out the appropriate referents of the IP at the subject's doxastic possibilities in terms of which the propositional concept is defined. The problem with this proposal is that if the MOPs are those used by the speaker, then this seems to make successful belief attribution too difficult. So the alternative proposal is that the reference-determining MOPs are those used by the subject instead.

But here is what is problematic with the latest proposal. On any reasonable account of the acquaintance relation or modes of presentation, it will be true that people can be acquainted with the same object in diversely different ways, employing very different MOPs to pick out the same thing. This would imply that the very same IP can express one proposition when used to attribute belief to one subject, but that it will most likely express a rather different proposition when used to attribute a belief to a different subject. This however brings us back to the problem faced by the modified version of Pietroski's proposal we looked at earlier, which is to account for our intuitions about

different people believing the same thing. For example, consider belief sentences with quantified NPs, as in :

(23) Every student believes that diamonds are made of silicon compounds

Since $IP_{23}$ is necessarily false, I assume that its intension is not what the CP denotes. It is hard to see how diagonalization can tell us what this proposition is though. For the students might use very different MOPs to pick out diamonds and silicon compounds, and have very different beliefs about the properties of these substances. If we define the diagonal proposition expressed by the IP using the MOPs of any one student, we will most likely end up with a proposition that is not believed by other students, even though they might all believe that diamonds are made of silicon compounds.

A similar problem comes up in explaining the validity of the following inference :

(24) Josep believed that Hersperus is Mars, and so did Suzanna

∴ Suzanna believed that Hersperus is Mars

Intuitively the above inference is valid, akin to the inference from "Josep hit Bill, and so did Suzanna" to "Suzanna hit Bill". The similarity of the two inferences is readily explicable if we take "believe" and "hit" to express two place-relations, and that in general, from "*X* ϕ-ed *Z* and so did *Y*" we can infer "*Y* ϕ-ed *Z*". So in order to account for the validity of (24), the IP of the premise and the conclusion have to express the same proposition. It is however difficult to see how this is possible on the diagonalization proposal. For Josep and Suzanna might use very different MOPs to pick out Venus and Mars. Most likely, the diagonal proposition expressed by the IP of the premise, defined

using Josep's MOPs, will be distinct from the diagonal proposition expressed by the IP of the conclusion, defined using Suzanna's MOPs. But then the inference is no more valid than inferring from "Josep hit Bill, and so did Suzanna" where "Bill" refers to Bill Clinton, to "Josep hit Bill", where "Bill" refers to Bill Corsby. If on the other hand the proposition expressed by both IPs is defined by the *speaker's* MOPs, we run the risk that although the same proposition is expressed, it is not one that either Josep or Suzanna believes in, because they are acquainted with Mars and Venus in ways very different from that of the speaker. So validity is preserved at the expense of soundness, which is equally unsatisfactory. Of course, it is perhaps possible that both the speaker and the subjects use the same MOPs to pick out the planets. But without a more detailed account of acquaintance and modes of presentation, we have no reason to think that this is a likely possibility. Furthermore, the validity of (24) should not depend on whether this empirical possibility obtains or not.

## 8. Interpreted Logical Forms

Many of the problems I have discerned for Pietroski's and Stalnaker's proposals are based on our intuitions about the correct truth-conditions of belief reports, and how to account for the validity of certain patterns of inference they enter into. Perhaps there are ways to get round these problems by denying the validity of some of such intuitions, or by introducing additional complexities in the semantics of CP. But I think a better strategy is available. The strategy I recommend is that what a CP denotes is to be identified with some quasi-linguistic object that is individuated more finely than sets of possible worlds. This does not mean the possible worlds theory of belief is to be given up though. Let me explain why.

The proposal I have in mind says that what a CP denotes is an *interpreted logical form* (ILF), a complex made up of a syntactic representation at the level of logical form together with the semantic values of the lexical items.[18] For our present purpose though, there is a simpler version available : the thesis is that a token of the CP "that *p*" in a context *C* denotes an ordered pair < *s* , *f* >, where *s* is the sentence type "*p*" that is the embedded IP, and *f* is an interpretation function which maps constituent expressions of *s* to the referents of their tokens in *C*. So for example, what $IP_{19}$ denotes is the ILF < "Hersperus is identical to Mars" , *F* >, where *F* maps "Hersperus" to Venus, "is identical to" to the identity relation, and "Mars" to Mars. Accepting such a proposal requires giving up the claim that a CP denotes a set of possible worlds. Nonetheless I think this proposal has four features that recommend itself to a possible worlds theorist.

First, ILFs are interpreted in that semantic values are assigned to the constituent syntactic expressions. It is therefore possible to give an account of their truth-conditions in terms of the syntactic structure of the constituent IP and the objects assigned by the interpretation function. This preserves our intuition that the objects of belief have truth-conditions and are bearers of truth and falsity.

Second, the individuation of ILFs is obviously very fine-grained. Unlike for example sets of possible worlds, necessarily equivalent ILFs need not be identical. This accommodates the very strong intuition that one can believe distinct but necessarily equivalent things. Furthermore, ILFs differ from for example structured meanings or

---

[18] See for example, James Higginbotham (1991) "Belief and Logical Form" in *Mind and Language* Vol. 6 No. 4, pp. 344-369; Gabriel Segal (1989) "A Preference for Sense and Reference" in *The Journal of Philosophy* , pp. 73-89; Richard Larson and Peter Ludlow (1993) "Interpreted Logical Forms" in *Synthese* 95, pp. 305-355. These proposals are not exactly the same however. In particular, Larson and Ludlow argue that to deal with attributions of demonstrative beliefs, ILFs should also include as constituents modes of presentation and particular events such as acts of pointing. I think this is a mistake but it does not bear on the main point in this section, namely that CPs denote something *more* fine-grained than sets of possible worlds.

Russellian propositions in containing syntactic items as constituents. Unlike these proposals then it easily take into account the observation that the truth-value of a belief report need not be preserved by substituting synonymous or coreferential expressions.

Third, on the present account, the referent of a CP does not depend on the mental state of the subject to which belief is attributed. This feature of *subject-independence* is not present in the modified version of both Pietroski's and Stalnaker's accounts. In Pietroski's case, the metalinguistic proposition that a CP denotes in a normal context is affected by the language spoken by the subject; with Stalnaker, the appeal to modes of presentation has the consequence that the diagonal proposition denoted by the CP is dependent on how the subject is acquainted with the objects mentioned in the IP. I have argued that this creates problems in accounting for people believing the same thing. On the other hand the ILF account avoids subject-dependence. The mental state of the subject of the belief report does not affect what the CP refers to, and so the problem does not arise.

Finally, the ILF proposal is a rather minimal *semantic* hypothesis that is appropriately neutral as between different *psychological* theories of belief. Although the proposal does say that it is ILFs and not sets of possible worlds that are among the *relata* of the believe relation, it says nothing about what this relation consists in. For example, it does not tell us whether only speakers of a language can have beliefs. Neither does it address the admittedly puzzling philosophical question of what is it to believe $p$ but not $q$ even though $p$ and $q$ are necessarily equivalent. It is open to the possible worlds theorist to claim that a psychological theory of what is it to believe an ILF involves relations to possible worlds. Such a theorist would indeed have to give up the semantic hypothesis that a CP denotes a set of possible worlds. But the ILF proposal does not introduce anything new in its ontology that a possible worlds theorist does not accept. Nominalist reservation

is thus not a reason for rejecting the proposal, and we have seen some of the positive reasons for endorsing it. A possible worlds theorist might claim, as before, that the belief state of a subject is defined by a set of doxastic possibilities, and that it determines which ILFs are believed by the subject. Such a theorist would have to provide an account that fills in the following schema :

(25) $< X , < s , f > >$ satisfies "$x$ believes $y$" in a context $C$
if and only if ____ $DOX(X)$ _____.

I think both Pietroski's and Stalnaker's proposals can readily be adapted to this schema. For example, Pietroski can claim that if $C$ is a normal context, then,

(26) $< X , < s , f > >$ satisfies "$x$ believes $y$" in $C$ if and only if the intension of $s$ includes DOX($X$) and, there is some sentence $t$ such that : (i) $X$ understands $t$ (in some appropriate way), (ii) $t$ is actually similar to $s$, (iii) $t$ is true at all possible worlds in DOX($X$).

On the other hand, Stalnaker can propose that,

(27) $< X , <s,f> >$ satisfies "$x$ believes $y$" in $C$ if and only if DOX($X$) is a subset of the diagonal of the propositional concept of $s$ in context $C$.

As suggested earlier, to spell out the proposal (27) in more details, one would have to say more about how the propositional concept is defined in terms of the contextually salient modes of presentation. As for Pietroski's proposal, as I have argued before, we need to know which feature of a context makes it normal or abnormal. We also need an explication of the relevant notion of linguistic understanding that is crucial to his theory.

Whether some plausible version of either account can be developed is a question that I shall set aside. The main aim of this paper is not to defend these theories, but to show two things. First, I have presented a dilemma for both Pietroski's and Stalnaker's proposals. On one hand their proposals in their original form suffer from certain serious problems in accounting for belief reports of non-English speakers. On the other hand plausible amendments lead to the undesirable consequence that the denotation of a CP is much too subject-dependent. It is hard to see how the modified proposals can account for (i) the validity of certain inferences that belief reports enter into, and (ii) the truth-conditions of belief reports with plural NPs. In the last part of this paper I suggest a way out. The idea is to distinguish between the *semantic* thesis that a CP denotes a set of possible worlds, and the *psychological* thesis that possible worlds are involved in a theory of the believe relation. The problem of subject-dependence can be avoided by giving up the semantic thesis. I point out the advantages of taking the referent of a CP to be an interpreted logical form, and how this is compatible with holding onto the psychological thesis. The proper evaluation of the psychological thesis is beyond the scope of this paper, but at least I hope what I have written here is relevant to how the debate might proceed.

# *Chapter Two*

# Concept Possession and The Language of Thought

Since the 70s, Jerry Fodor has been arguing that any adequte theory of cognition will have to postulate a language of thought. Although Fodor himself regards this as truth beyond doubt, the thesis remains an empirical hypothesis about the nature of mental representation. So it is surprising to see that an *a priori* argument for the language of thought has recently been advanced by Martin Davies.[1] Building upon certain ideas from Gareth Evans, Davies argues that the process of making inferences is one where the possession of concepts play a "causally systematic" role. Furthermore, it is supposed to follow from this fact that those mental states which enter into such a process have syntactic structure. In this paper I shall show that the argument fails because it equivocates between a strong and a weak sense of "causal systematicity". Moreover, the argument is unsound on either the strong or the weak interpretation.

## 1. Why should there be a Language of Thought?

For the purpose of this essay I shall take the language of thought hypothesis (LOT) as the thesis that mental representations have a combinatorial syntax. That is, there are mental representations which are complex and are composed of other representations. By mental representations I mean the representational states or objects which are postulated

---

[1] Martin Davies (1991) "Concepts, Connectionism, and Language of Thought" in Ramsey, Stich and Rumelhart (eds.) *Philosophy and Connectionist Theory* New Jersey: Lawrence Erlbaum Associates. Page references are given in square brackets.

by psychological theories. So LOT is plausible to the extent that it is part of our successful psychological theories. There are general arguments that aim to show just that, that successful psychology will most likely be committed to LOT.[2] It would be useful to look at two such arguments to contrast them with Davies's new argument.

The first one, the *productivity* argument, has it that at any given time, there is no limit on the number of thoughts we can think in principle. But how is that possible? An obvious explanation is that thinking involves mental representations, and our mental representations have a combinatorial syntax such that there is no limit on the number of complex representations there are. Hence there are indefinitely many thoughts we can have. This argument, however, rests on the substantial productivity assumption that there are infinitely many thoughts we can have *in principle*. Such an assumption is not uncontroversial. Opponents of LOT might agree that there are good reasons for adopting the assumption as a *methodological* principle (for example, it precludes hypothesizing that a system solves a computation problem by storing all possible solutions). Nonetheless they might insist that the assumption is strictly speaking false, and that our psychological capacities are essentially finite.

The *systematicity* argument for LOT, however, is one that does not assume productivity. It is claimed that whether or not our psychological capacities are unbounded, it is at least the case that the relation of systematicity holds among the thoughts we actually can have. The idea is that if we can think a thought with certain conceptual components, then we can think a semantically related thought where the components are arranged differently eg. as with the pairs of thoughts *John hit Mary*, and *Mary hit John*.

---

[2] See for example Fodor and Pylyshyn (1988) "Connectionism and Cognitive Architecture" in Pinker and Mehler (ed.) *Connections and Symbols* Cambridge : MIT Press, pp. 3-71.

Systematicity is supposed to be a reason for accepting LOT because this feature of our thoughts can easily be explained given the hypothesis.

How might the opponents to LOT respond to the systematicity argument? First they might suggest that there are alternative hypothesis which can also explain systematicity without assuming LOT. They might also argue that there are psychological phenomena beside systematicity which are better explained by the rival hypothesis, so that LOT is not more plausible overall. I think this is the position of many connectionists who are opposed to LOT. They see themselves as engaging in a developing research project which has no need of the productivity assumption, and which can explain systematicity without assuming LOT. At present, whether this is possible remains to be seen. But the point is that against such aspiring researchers, the productivity and systematicity arguments will inevitably appear question-begging.

This is where Martin Davies's *a priori* argument comes in. Davies does not argue for LOT by saying that it leads to successful research or good explanations. He begins by agreeing with Evans that the so-called "generality constraint" is an *a priori* principle that thinkers have to satisfy. The admittedly intuitive and plausible idea behind this principle is that thinking is systematic : to have a thought requires the possession of conceptual capacities, and these capacities can be exercised in conjunction with each other to enable the subject to entertain a range of different thoughts. This is of course rather similar to the idea of systematicity just discussed. What is intriguing about Davies's argument is that LOT is supposed to be a *consequence* of this fact. The argument is not that, among various competing explanations, LOT provides the best empirical theory of how there can be thinkers that satisfy the generality constraint. Rather, the claim is that *necessarily* only creatures for which LOT is true can have the systematic capacity for thoughts and thus satisfy the generality constraint.

If this argument is sound, it would give us an extremely powerful conclusion indeed. For we can now conclude, *prior* to the development of connectionist research, that such theories cannot be correct theories of thinking unless they implement a language of thought. One might naturally be suspicious of how such an argument is possible. How is it possible to establish an empirical claim on the basis of *a priori* considerations? I think such worries are legitimate, and the thrust of this paper is that indeed we have no *a priori* reason to accept LOT other than empirical ones. But this requires looking at the details of the argument. One might argue that in the analysis of a certain concept, we find that the concept applies only if certain empirical conditions obtain. I think this is what Grice took himself to be doing when he defended the causal theory of perception, that our concept of the perceptual relation is one that obtains only if some appropriate causal chain is present.[3] We might see Davies as pursuing a similar project, arguing that the concept of *a thinking subject* applies only under certain conditions from which LOT follows.

## 2. Davies's notion of syntax

Still, one might naturally wonder how our concept of a thinking subject can possibly commit us to a contentious thesis such as LOT. One way to allay the suspicion is to point out that the LOT thesis being defended is extremely weak. First of all, the thesis concerns only those thoughts that enter into inferential thinking, mental states which we can be conscious of. The argument has nothing to say, however, as to the nature of those mental states to which we have no conscious access.[4] Furthermore, not only does Davies's

---

[3] See Paul Grice (1989) "The Causal Theory of Perception" in *Studies in the Way of Words* Cambridge : Harvard University Press, esp. page 224.

[4] This is a point that Ned Block has raised regarding the limitation of the systematicity and productivity arguments. See his paper "The Computer Model of the Mind" in Osherson and Smith (1990) (eds.) *Thinking : An Invitation to Cognitive Science,* Volume 3, Cambridge : MIT Press.

argument applies to a restricted class of mental states, the conditions that these states have to satisfy in order to possess syntactic structure is also very weak. According to Davies, a representational state or object has syntactic structure when the content of the state is systematically dependent on the syntactic properties it has. A syntactic property is any property that satisfies these three conditions :

(S1) It is a physical property.

(S2) It "correlates" with some semantic property *S* (this is relative to a system or subsystem, so *P* correlates with *S* in a (sub)system if and only if anything in the (sub)system has *P* if and only if it has *S*).

(S3) It is a "determinant of causal consequences". That is, it is a causally salient property of the state that has it.

I shall call Davies's notion of syntactic structure *minimal syntax*. Minimal syntax is minimal indeed, for I think just about every kind of representations that theorists have proposed can possess minimal syntax. First there are what we might call linguistic symbols, symbols which are generated by a set of recursive rules and a set of primitive expressions. Obviously such symbols can possess minimal syntax. However, there also are map or picture like representations for which it is not so clear what the rules or the primitive expressions are. Nonetheless on Davies's account these can posess minimal syntax too. For example, consider a system that uses a map as internal representation. We can imagine that in such a system, the property of having a red dot on the map correlates with representing a city. Both the causal role of the map and its contents (e.g. that there are three cities in a certain country) are systematically dependent on whether such physical properties are instantiated. Such a representation can satisfy conditions (S1) to (S3) and so possesses minimal syntax.

What is common to sentences, maps or pictures is that they possess what we might call *proper syntax* : (tokens of) complex representations have either spatial or temporal *proper parts* that are themselves representations. I think most of what we take to be structured representations do indeed have proper syntax, including computer data structures.[5] But minimal syntax is even weaker than proper syntax. To see that there can be minimal but not proper syntax, we can consider this toy example that Davies discusses. There is a drinks machine that accepts coins to produce tea or coffee. The coins are either round or square in shape, and red or blue in colour. The machine produces a coffee as output if the input is square, and tea instead if it is round. Milk is added if the input is red, but not if it is blue. Davies suggests that we take *squareness* as a syntactic property, correlating with the semantic property of meaning something about coffee. Similarly, *redness* correlates with meaning something about milk, etc.. A coin that is red and square might then be taken to be a structured input that means *the client wants a coffee with milk*, the content of the coin being a function of the contents of those syntactic properties it has. According to Davies, this is a case where the inputs are syntactically structured, since their contents are systematically dependent on their syntactic properties - "the formal language of its input states has just four primitive symbols and one binary operation" [239]. Presumably the suggestion is that the properties of *redness*, *squareness*, *blueness*, *roundness* are taken to be the symbol *types*, and property conjunction is the operation that takes *squareness* and *redness* say into the complex symbol *being square and*

---

[5] Fodor and Pylyshyn disagree. They write (*op. cit.*, page 57), "there is no necessity that a token of an atomic symbol be assigned a smaller region in space than a token of a complex symbol; **even a token of a complex symbol of which it is a constituent** ... functional elements can be physically distributed or localized to any extent whatever. In a VAX ... pairs of symbols may certainly be functionally adjacent, but **the symbol tokens are nevertheless spatially spread through many locations in physical memory**." It seems to me that they have confused two issues here : (1) the relative sizes of a complex symbol and its symbol parts; (2) how might a symbol spread out in space. While it is true that a string such as "holism is bad for your health" might be stored across different (possibly spatially scattered) memory locations in the VAX, it is false that there are any parts of the string which occupies a *larger* region of space than the whole string. The proper parts of a scattered objects do occupy a smaller spatial or temporal region than the whole.

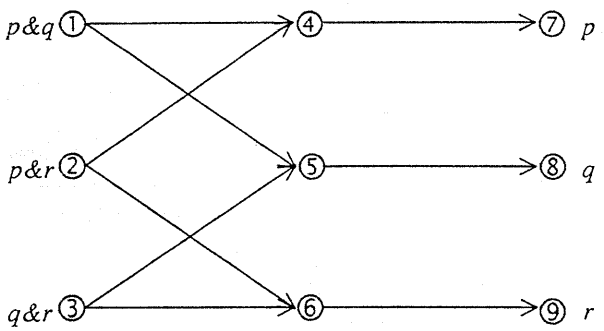
*red*. An input coin being a token of a complex symbol type is then a matter of instantiating the property that *is* the type. Note that the square red and square blue inputs share a common syntactic constituent on Davies's account, but this is not in virtue of having proper spatial or temporal parts that are of the same syntactic type. So this is a case where we have minimal but not proper syntax.

But if properties are admissible as symbol types, are there any restrictions on *which* kind of properties they have to be? In discussing what syntactic properties are, Davies briefly raises the question of whether they have to be intrinsic or not, but seems to have left the issue open. In his own example *redness* is taken to be a symbol type, but of course colour properties are often regarded as secondary properties and hence relational. So perhaps Davies is willing to allow relational syntactic properties. But are there any criteria as to *which* relational properties are admissible as syntactic ones? Davies does not say, but I think whether his argument provides a serious critique of connectionism depends very much on how this issue is settled.

Let us then consider how connectionist representations relate to Davies's notion of syntactic structure. A typical connectionist network consists of layers of computing units. A unit can excite or inhibit other units through links with adjustable characteristics. A *localist* network is one where all representations consist of individual units. (They are all *local* representations.) In the current debate between classical theories of cognition (which all postulate a language of thought) and connectionism, it is generally considered that localist networks do not have structured representations. There might be semantic and causal relations between individual units, but I think most theorists agree that there are no part/whole relations that hold between units and that all localist representations are atomic. So for example when Fodor and Pylyshyn argue that networks cannot explain

cognition because they do not have structured representations, Smolensky readily concedes that their critique applies to localist networks.[6]

However, on Davies's notion of minimal syntax, even a localist network might turn out to have structured representations. Consider for example the following network :




The diagram above depicts a network of nine units. Units 1 to 3 are the input units and units 7 to 9 are the output ones. When any input unit is "activated", they in turn cause other units to be activated with the direction of activation indicated by the arrows. Thus for example, if unit 1 is activated then it in turn activates units 4 and 5 (but not 6), leading to the activation of units 7 and 8. The content of the input and output units are written next to them. (Imagine the letters being replaced by sentences.) We might say that the network performs deductive inferences in the sense that its input-output transitions are truth-preserving. As pointed out earlier, the individual units in such localist networks are generally regarded as unstructured. But this is not so on Davies's notion of syntax. Given that relational properties can be symbol types, there is no reason why we cannot take the input units of the network as composed of the following atomic symbols :

---

[6] See Smolensky (1988) "Connectionism, Constituency and the Language of Thought" in Loewer and Rey (eds.) *Meaning in Mind : Fodor and his critics* Oxford : Blackwell.

property $F$ = the property of *having an activation link to unit 4*

property $G$ = the property of *having an activation link to unit 5*

property $H$ = the property of *having an activation link to unit 6*

These three properties certainly satisfy the conditions (S1) to (S3). They are first of all physical properties. Each is correlated with a different semantic property : an input has $F$ if and only if it means something about $p$, etc.. The three are also causally relevant properties which determine how an input unit might activate other units in the network. As in the drinks machine, we can assign content to these atomic properties : $p$ to $F$, $q$ to $G$, and $r$ to $H$. Each input unit is thus "composed" of two atomic syntactic properties - $F$ and $G$ for unit 1, $F$ and $H$ for unit 2, $G$ and $H$ for unit 3. Finally, the contents of the units are systematically dependent on these properties. So the input units of the networks do possess minimal synatx.

Such a weak notion of syntactic structure is certainly non-standard. But "syntactic structure" being a technical term, this is by itself no reason for criticism. The question however, is whether the class of structured representations so construed forms an interesting class. If not, this would diminish the theoretical interest of Davies's argument. It is then not really clear what his argument *for* LOT is arguing *against*. But as the following passage suggests, Davies seems to think that his LOT thesis is incompatible with a certain type of non-local representations. If this is right, then his argument would then establish a substantial and interesting conclusion, since it allows us to reject the adequacy of a class of networks that employ only such representations. According to Davies [247],

> typical connectionist networks do not exhibit causal systematicity of process, and syntactic structure in input states. Of course, connectionism

> comes in several varieties, and there are some networks that do have these features; examples can be provided by networks with local representation of all the primitive concepts of some classical task analysis. So let us be more specific. What is to be considered is connectionism with distributed representation. In particular, we focus on networks with microfeatural, dimension shifted, representation in the style of Smolensky.

Distributed representations are those where content is assigned to patterns of activation over a group of units. As a simple example we might consider a group of four units representing respectively *coffee*, *tea*, *is hot*, *is cold*. The simultaneous activation of the *coffee* and *is hot* units might be taken to mean *coffee is hot*, and the activation of the *coffee* and *is cold* units might be taken to mean *coffee is cold*.[7] In this simple case, all the patterns that represent something about coffee will have a common subpattern of activation - the activation of the *coffee* unit. On Davies's account this is a straighforward case of structured representation. Davies's target however is Smolensky, who favours distributed representations of a rather different sort. With Smolensky-style representations, although the different patterns of activation over the same group of units *G* can have semantically related meanings, eg. representing *coffee in a cup*, *coffee in a bowl*, etc., there is no single subpattern of activation that occurs if and only if the total pattern means something about coffee. There might be subpatterns that *resemble* each other in some way, but "there simply is no strictly common subpattern of activation that can be identified as a syntactic element meaning coffee." Davies concludes that "If Smolensky is right then, relative to a semantic description in terms of coffee, cups, jugs,

---

[7] This might be what Davies has in mind when he says in the quoted passage that there is syntactic structure if there is local representation of "all the primitive concepts of some classical task analysis". In the simple example we have both local representations of concepts such as *coffee* and structured representations of propositional contents. But if Davies thinks that local representations of concepts *entails* structured representations then again I think most theorists would disagree. For it is quite possible to have a local network where all contents - propositional or not - are represented by local, atomic representations.

and the like, the input states of a typical connectionist network with distributed representation will not have a syntactic description." [248-9]

It seems to me that this last claim is too strong. A subpattern of activation is by no means the only candidate for a syntactic element, on Davies's account, since it allows syntactic constituents to be *relational* properties. Even though all the activation patterns that mean something about coffee do not share a unique subpattern, they might still count as structured representations if they share an appropriate relational property that is not possessed by other patterns. e.g. it might be that all and only activation patterns that mean something about coffee have the property of *causing activation of such and such units further down the road.* Surely this is a causally salient physical propoerty. After all, the causal role of an activation pattern over a group of units *G* depends on both (a) *which* units of *G* have been activated, and (b) *how* the units in *G* are wired up to other units. Finally, over that group of units, this relational property correlates with the semantic property of *meaning something about coffee*. So all three conditions on syntactic property have been satisfied. If similar relational properties can be found to correlate with other semantic properties, there is no reason why we cannot assign minimal syntax to activation patterns that mean *coffee in a cup, coffee in a bowl*, etc..

I am not claiming that if cognition can be explained in terms of connectionist networks employing such distributed representations, then they *must* turn out to have minimal syntax. But the examples show that they *might*. Here I just want to point out how extremely weak the notion of minimal syntax is. It turns out that structured representations can include sentences, maps, pictures, both local and distributed connectionist representations, and probably many others. One might wonder then, what is the *scientific* interest of Davies's argument. After all, the requirement that thoughts must have minimal syntax does not seem to provide any interesting guidance as to how

psychological theories are to be constructed. Having said that, it has to be pointed out that Davies does take himself to be defending only a "minimal" version of LOT. Perhaps the *philosophical* interest of the argument is to show that there is some way to understand the LOT hypothesis as a tenable, non-contentious thesis. This would also be good news for philosophers such as for example Hartry Field, who has argued that the problem of intentionality can only be solved by postulating a system of structured mental representations.[8] Unfortunately, I think Davies's argument is not valid even given such a weak construal of the LOT thesis. Let me now explain why.

## 3. From Causal Systematicity to Syntax

According to Davies, inferential thinking is a mental process that has thoughts as inputs and outputs. So given for example the thought *Joe is a bachelor* as an input, the inferential process might produce *Joe is unmarried* as an output. The conclusion that Davies seeks to establish is that the thoughts that are the inputs to such a process have minimal syntax. Nothing is said as to the structure of the outputs. But if those very thoughts that are the outputs can also appear as inputs to further inferences, then presumably the outputs would also have to be structured. In any case, our focus is on the argument for structured inputs, which comes in two steps. First, it is argued that if any input/output (I/O) process is "causally systematic" relative to semantic patterns, then the inputs to such a process possess syntactic structure. Next, the argument is that inferential thinking is precisely such a process. Since thoughts are the inputs to inferential thinking, the conclusion is that thoughts possess syntactic structure. Thus LOT is true. The argument is obviously valid, and so the question is whether we should accept either premises. We will start by considering the first.

---

[8] See Field (1978) "Mental Representation" in *Erkenntis* 13, No. 1, pp. 9-61.

Davies's argument makes crucial use of the idea of *causal systematicity*, which he introduces as follows :

> Suppose a generalization G describes a pattern to be found in the input-output relation of some physical system. If we consider several input-output pairs that exhibit the common pattern, then we can ask whether the several input-ouput transitions have a common causal explanation corresponding to the common pattern they instantiate. If there is a common causal explanation, then we can say that the process leading from those input states to output states is causally systematic relative to the pattern described by G. [233]

One problem with this definition is that whether two I/O transitions have a common causal explanation will in general depend on the level of explanation : on the *microphysical* level for example, the transitions probably involve very different configurations of particles acted on by forces of different magnitudes, etc.. In that case there will not be a common causal explanation of why those transitions conform to the pattern. But of course this does not preclude a common explanation at a higher level. Davies seems to recognize this, and he indicates later on that what causal systematicity requires is a common *mechanism* that mediates the transitions conforming to the pattern :

> if we think of a physical system as containing various subsystems or mechanisms, then the requirement for causal systematicity relative to the pattern described by G is that **there should be a mechanism whose presence in the system explains all the input-output transitions that conform to the pattern described by G.** [235]

Davies however does not say how we are to individuate mechanisms. So are we allowed to take the inferential system *as a whole* to be a mechanism? Surely *its* presence will explain *all* I/O patterns, including G! Elsewhere, in discussing whether connectionist nets exhibit

causal systematicity, Davies implicitly assumes that the mechanism should only be *part* of the I/O system, and mediates *exclusively* those transitions that instantiate the pattern in question :

> To ask whether the process that is going on in the network is causally systematic relative to that pattern is to ask whether the *coffee* to *warm drink* transitions all have a common explanation; whether there is, as a component of the network, a mechanism that is responsible for **all and only those transitions**. [249]

So in the case of the drinks machine, suppose there is a mechanism which can recognize if an input is square. If it detects a square input, it makes a cup of coffee, and otherwise it does nothing. We might call this mechanism COFFEE, and the existence of COFFEE would imply that the machine is causally systematic relative to the pattern "square input → coffee output". If there is no such autonomous mechanism that mediates all and only the transitions that conform to the pattern, then the pattern would not be causally systematic on Davies's criterion. Notice that causal systematicity is a very strong requirement. We will have reason to doubt whether inferential thinking is indeed causally systematic.

Now that we have clarified somewhat the notion of causal systematicity, let us see how it relates to syntactic structure. The first point to note is that a causal systematic pattern entails that the inputs have some *triggering property* that correlate with the property specified in the pattern, in the following sense. If a system exhibits a causally systematic pattern "$F$ input → $K$ output", then by definition there is a mechanism that is exclusively responsible for producing $K$s upon inputs that are $F$s. There must then be a causally salient property - a triggering property - that all and only the $F$s have, and which explains why they alone can activate that mechanism to produce a $K$. If the process has a

variety of different I/O patterns which are causally systematic, then there would be many more triggering properties that correlate with the properties cited in the patterns.

Consider for example Davies's machine. If we know that mechanism COFFEE mediates every instances of the pattern "square input → coffee output", and does not mediate instances of other patterns, then we know that there has to be some triggering property unique to the square inputs, a property that the round inputs do not have, such that all and only the squares can activate COFFEE. (Of course, we cannot conclude that COFFEE recognizes the squares by their shape, for it is possible that only the squares are conductors, and COFFEE detects them by passing a current through them.) This conclusion does not follow if the pattern is not causally systematic. For suppose instead of COFFEE we have two totally autonomous mechanisms B-COFFEE and W-COFFEE. B-COFFEE only makes black coffee, and W-COFFEE makes only white coffee. Because the pattern "square input → coffee output" is not causally systematic, we cannot conclude that the square inputs share a common triggering property. The relevant crucial point, which I think is correct, is that where we have a causally systematic pattern "*F* inputs → *G* outputs", there is a causally salient property *C* of the inputs that correlates with property *F*.

Davies then argues that if the causally systematic patterns are *semantic* ones, it follows that the inputs have causally salient, physical properties that correlate with their semantic properties. But since all three conditions for minimal syntax are satisfied, it also follows that the inputs to the process have syntactic structure. Again Davies offers his drinks machine as an illustration. Here are the inputs, their meanings, and their corresponding outputs :

| input | meaning | output |
|---|---|---|
| square blue | the client wants coffee without milk | coffee without milk |
| square red | the client wants coffee with milk | coffee with milk |
| round red | the client wants tea with milk | tea with milk |
| round blue | the client wants tea without milk | tea without milk |

So one semantic pattern of the I/O process is that coffee is produced whenever the input means that the client wants coffee (with or without milk). Suppose it is the mechanism COFFEE that mediates instances of this pattern. So there is some causally salient property *C* ( e.g. being a conductor ) in virtue of which the square inputs activate COFFEE. Now suppose also we have this dedicated mechanism MILK which adds milk when and only when there is a red input. So red inputs must share a different salient property, say *M*, in virtue of which they activate MILK. Now *M* correlates with the semantic property of meaning that the client wants a drink with milk, and *C* correlates with meaning that the client wants coffee. So these are syntactic properties according to Davies's minimal syntax. A red and square input has both *M* and *C*, and its meaning is a function of the semantic properties that *M* and *C* correlate with. We can therefore take the syntactic properties *M* and *C* to be the syntactic constitutents of the input. We might contrast this with a case where the machine is made up of four distinct and autonomous mechanisms : one that makes white coffee, one that makes black coffee, etc.. The semantic pattern "input means that client wants coffee → coffee is produced" is no longer a causally systematic one, and we can no longer conclude that a red square input has two syntactic properties as its syntactic constituents.

I think we can accept Davies's claim that minimal syntax follows from causal systematicity with respect to semantic patterns. However, it is important to note that the

truth of this claim rests on two crucial assumptions, namely that (a) the semantic patterns have to be *strict*, and that (b) the semantic patterns have to be mediated by *dedicated* mechanisms. Give up either assumption, and the argument would collapse. To see why this is the case, consider for example this pattern (G) :

(G) If an input to the system has semantic property *P*, then the system produces an output with property *Q*.

Davies wants to argue that if (G) is causally systematic, then there is a causally salient property *C* that correlates with the semantic property *P*. That is, the inputs to the system has *C* *if and only if* it has *P*. My claim is that first, to show that *all* *P*-inputs have *C*, pattern (G) has to be strict. Second, to show that *only* the *P*-inputs have *C*, the mechanism that mediates (G) has to be dedicated. Here is why.

What I mean by a strict pattern is one that has no exceptions under normal conditions : all the input-output transitions should conform to the pattern if the system is operating normally, with no malfunction, human or divine interference, etc.. So if (G) is a strict pattern, this means that normally every input that has *P* is followed by an output that has *Q*. It is of course compatible with (G) being strict that there are inputs to the system that do not have property *P*. But if even under normal operations there are *P*-inputs that do not lead to *Q*-outputs, then (G) is not a strict pattern. In this case, even if *some* of the *P*s always manage to activate a particular mechanism *M* to produce a *Q*, and do so in virtue of having some triggering property, we have no reason to conclude that this triggering property is shared by *all* the *P*s. So there might not be a syntactic property that correlates with *P*.

But not only does pattern (G) have to be strict, the mechanism that mediates transitions conforming to the pattern has also got to be a dedicated mechanism, one that *exclusively* mediates transitions conforming to the pattern. That is, the mechanism is one that is activated only by inputs with property *P*. This requirement is necessary because if there are other inputs that activate the mechanism, but which lacks the semantic property *P*, we cannot then conclude that there is a triggering property possessed by *only* the *P*s.

The first premise of Davies's argument for LOT is that a system has inputs with minimal syntax if the I/O semantic patterns are causally systematic. We now see that this is true only if these semantic patterns are strict patterns mediated by dedicated mechanisms. Let us call this notion of causal systematicity "*strong* causal systematicity", to be distinguished from "*weak* causal systematicity" where non-strict semantic patterns are mediated by non-dedicated mechanisms. (There are of course other intermediate "mixed" notions, but they need not concern us here.) The crucial point is that the first premise is true only if we understand causal systematicity in the strong sense. The strong sense of causal systematicity does seem to be what Davies has in mind when he argues for his first premise. When he considers whether networks exhibit causal systematicity he does say that this requires "a mechanism that is responsible for **all and only** those transitions" [249]. His own examples of causally systematic processes, such as those of the drinks machine, are also cases where strong systematicity obtains. So, to complete his *a priori* argument for LOT, Davies needs to show that inferential thinking is causally systematic in the strong sense. However, I think this is something that he fails to argue for. As we are about to see, what Davies shows at most is that inferential thinking is causally systematic in the weak sense. But then the argument for LOT is not valid because it rests on an equivocation between strong and weak causal systematicity. If we interpret

the second premise in the strong sense however, we have every reason to believe that it is false.

## 4. Conceptual Capacities and Inferences

Davies's argument for his second premise assumes what we might call the "neo-Fregean" theory of thoughts, which derives from some of Gareth Evans's ideas. For our present purpose the two main relevant claims are :

(a) Having thoughts requires having concepts.

(b) Having concepts is to be construed in a "full blooded" way.

On the neo-Fregean theory, the content of a thought is composed of concepts, and to have a thought one must grasp or possess those concepts that are the constituents of its content. So for example to be able to think that *a is F*, say, the subject has to have the concept of *a* and the concept of *F*. In general, to have the concept of $\phi$ is to know what it is for something to be $\phi$. According to Evans, having concepts is a matter of having conceptual capacities that can be exercised in conjunction with each other. Thus it is supposed to be a conceptual truth that if a subject can have the thoughts *a is F* and *b is G*, then the subject must also be able to think *a is G* and *b is F*. This is what is known as the "generality constraint". What is important for Davies's argument, is that the attribution of concepts to a subject is not simply a claim about what thoughts the subject is capable of having if he were able to entertain certain other thoughts. According to Evans and Davies, the state of having a certain concept is a genuine mental state of the subject that has a causal role. It enters into the causal explanation of the inferences that the subject carries out. This is what is means to construe the possession of concepts in a "full-blooded" way. The causal

systematicity of inferential thinking is supposed to be a straightforward consequence of this fact :

> A thinker who has the thought that *a is F* appreciates that from this thought it follows that *a is H*, say; and he also appreciates that from the thought that *b is F* it follows that *b is H*. ... The two inferences are manifestations of a common underlying capacity; namely, mastery of the concept of being *F*.
>
> As Evans himself makes clear, the notion of a capacity or dispositions is not to be understood in terms of the bare truth of conditional statements, but rather in a "full-blooded" way. The idea of a common capacity being manifested in the two inferences should be unpacked in terms of a common explanation, adverting to a common state. In short, there is causal systematicity relative to the input-output pattern in a thinker's inferential practice. [244]

I must confess that I find the argument extremely cryptic to say the least. It is one thing to claim, as Evans does, that a particular conceptual capacity is causally relevant to inferential transitions that conform to some (possibly non-strict) semantic pattern. It is surely quite a different claim to say that semantic patterns in inferences are strict, and that they are mediated by dedicated mechanisms!

Davies does go on to offer an example, and appeals to the philosopher's favourite concept of a bachelor. The idea is that if a subject *S* infers from *Bruce is a bachelor* to *Bruce is unmarried*, and also from *Nigel is a bachelor* to *Nigel is unmarried*, then it is supposed to be the case that his ability to make these inferences "depends in each case on the same general capacity, namely the mastery of the concept of being a bachelor." [244] So suppose we accept that,

(1) Subject *S* performs inferences that conform to the pattern (P) : "*S* infers from *x is a bachelor* to *x is unmarried*"

(2) The state of having the concept of a bachelor is causally relevant in those of *S*'s inferences that conform to pattern (P).

Why should strong causal systematicity follow from (1) and (2)? *Davies simply does not say.* The state of having a certain concept is some particular mental state alright, but a state is surely not a mechanism. So even if we accept both (1) and (2), what reason do we have for thinking that there is some dedicated mechanism that mediates (P)? Perhaps the implicit assumption is this :

(3) If the state of having a certain concept *C* is causally relevant to inferences that conform to a semantic pattern, then there is a corresponding mechanism *M*(*C*) that operates in all I/O transitions that conform to the pattern.

Perhaps the thought is that given (1) to (3) we can then conclude that the inferential process is causally systematic with respect to semantic pattern (P). I think we certainly have no *a priori* reason to accept (3), and I shall come back to this point later. But even if it were true, this only shows that *weak* causal systematicity obtains, that there is a mechanism *M* that operates in all transitions conforming to (P). If Davies indeed has weak causal systematicity in mind when he defends his second premise, then his argument for LOT would simply be fallacious because it equivocates between the strong notion in the first premise and the weak notion in the second. This perhaps explains why in defending the second premise Davies fails to do two things : (i) show that (P) is a strict semantic pattern, and (ii) explain why the presence of a state causally relevant to transitions conforming to (P) entails that there is a dedicated mechanism involved.

Be that as it may, do we have any reason to accept Davies's argument for LOT on a non-equivocating reading? Since we saw that the first premise is true only if we understand causal systematicity in the strong sense, let us also take the strong reading of the second premise. So should we accept that for example pattern (P) is a strict semantic pattern mediated by a dedicated mechanism? Obviously no. First, strict semantic patterns are hard to come by, and (P) certainly is the least likely candidate. From *x is a bachelor*, a subject might infer, not *x is unmarried*, but instead : *x is an unmarried man*, *x is a man*, *x is not a rock*, etc.. These are all cases of deductive inferences that are exceptions to (P), and the conditions under which a subject carries out such inferences need not be in any sense abnormal at all. Furthermore, Davies has given no reason for thinking that the I/O system that carries out deductive inference might not also be involved in non-deductive inference. But if that is the case, the absence of semantic patterns in inferential thinking is even more overwhelming. From *x is a bachelor*, depending on one's store of beliefs and who knows what else, the subject might practically infer anything. But if the semantic pattern (P) is not strict, then as we saw earlier, even if there is a mechanism that takes some of the *bachelor* inputs to produce *unmarried* outputs, there need not be a causally relevant syntactic property that is common to *all* inputs that mean *x is a bachelor*.

It might perhaps be replied that *P* is not a very good candidate for a strict semantic pattern. But surely "input means *x is a bachelor* → output means something" is one? I grant that this pattern is strict, but now the problem is what *a priori* reason we have for thinking that there is a special state that is causally relevant to all inferential transitions conforming to this pattern. One instance of this pattern is the inference from *Joe is a bachelor* to *Joe exists*. Is it *a priori* obvious that the state of having the concept of a bachelor might still be causally relevant in this particular transition? What about inferring from *Joe is a bachelor* to *Joe is a bachelor or snow is white*? Furthermore, even

if we were able to find a strict semantic pattern whose instances always invoke the state of having some concept *C*, the most that we can conclude is that weak causal systematicity obtains. It still does not follow that the mechanism corresponding to that concept is dedicated to mediating the pattern. As an example, suppose it is true of Fred that (P) is a strict semantic pattern, that as a matter of fact Fred always infer from *x is a bachelor* to *x is unmarried* and nothing else. We might agree with Davies that the state of having the concept of a bachelor is causally relevant in all these inferences. But what if Fred also infers from *x is married* to *x is not a bachelor*? Surely the state of having the concept of bachelor is also causally relevant? But then if there is a mechanism that corresponds to such a state, it is not one that is dedicated to (P). So strong causal systematicity still fails.

Later on, Davies speaks of a piece of knowledge which is partially constitutive of the state of having a certain concept e.g. part of what is it to have the concept of a bachelor is to have the knowledge that if someone is a bachelor, then that person is unmarried. This piece of conceptual knowledge is said to be "implicated" in inferences from *x is a bachelor* to *x is unmarried*. Of course, we have no more reason now than before that strong causal systematicity obtains. For this piece of knowledge is presumably also "implicated" in inferences from *x is married* to *x is not a bachelor*, and also in some cases of English speakers, in metalinguistic inferences from "___ is a bachelor" *is true* to "___ is unmarried" *is true*. No *a priori* conclusion can be made as to whether there is some mediating mechanism exclusive to inferences of any semantic pattern.

In fact, it seems a misguided attempt to think that we can gain deep insights into the nature of inferential mechanisms by reflecting upon concept possession or conceptual knowledge. Here we might employ David Marr's distinction of the three levels on which an I/O process can be understood. A description of the first level states which function is computed by the I/O process; the second level describes the algorithm which computes the

function and the representations employed by the algorithm; the third level is the level of hardware implementation, how the algorithm is realized by the mechanisms that carry out the I/O process. Davies's notion of causal systematicity concerns the nature of the mechanisms involved in an I/O process, and so belongs properly to the third level of hardware. Where do we place explanations that cite knowledge? Christopher Peacocke has argued for the existence of an intermediate "level 1.5" that describes the information "drawn upon" by an algorithm. Such a level is so labelled because on one hand different algorithms can draw upon the same information, and on the other hand the same function can be computed by algorithms that draw on different information.[9]

I think Davies's talk of conceptual capacities and conceptual knowledge belongs to level 1.5 - to say that the knowledge of bachelors being unmarried is causally relevant to inferences from *x is a bachelor* to *x is unmarried* is to say that the inferential process draws upon the information that bachelors are unmarried ( of course an inference can also draw upon misinformation ). As Peacocke points out, the notion of information being drawn upon is a causal one :

> It requires that a state which carries the information drawn upon is causally influential in the operation of the algorithm or mechanism; indeed it requires that the algorithm or mechanism produce states with the content they do in part because of the content of the information-carrying state.

Accepting this requirement would indeed respect Evan's insistence that conceptual capacities are to be taken in a "full-blooded" way, for it requires the existence of states which contain information pertaining to the concepts, and which are causally relevant in inferences that the subject is disposed to make. However, the claim that the same

---

[9] Christopher Peacocke (1986) "Explanation in Computational Psychology: Language, Perception and Level 1.5" in *Mind and Language* Vol.1, No.2, pp. 101-123.

information is drawn upon in different inferential transitions says nothing about whether the same algorithm or mechanism is involved. Neither does it preclude the same information from being drawn upon in transitions of different semantic patterns, or that a certain inferential transition might draw on more than one piece of information. Furthermore, it is an open empirical question to what extent such information might also be available to other mental processes which do not have thoughts as inputs, as in the fixation of perceptual beliefs. Whether there is causal systematicity with respect to semantic patterns is an issue that belongs properly to level 3 - the level of hardware implementation. The causal relevance of conceptual capacities does not answer the question one way or another.

This suggests that we might construe the notion of causal systematicity as describing an I/O system at level 1.5. To say that a pattern of inference is causally systematic, is to say that there is a piece of information that is drawn upon in all inferences of that pattern. There is no commitment to the pattern being strict, or that the information-bearing state cannot also be causally relevant in other inferential patterns. If we understand causal systematicity this way, what can we conclude *a priori* about the syntactic structure of thoughts? Not much, I think, given that we often do not have conscious access to the intermediate stages of our inferences. Furthermore, the question of whether two inferences draw upon a common piece of information does not seem to be an issue that can be settled *a priori*. Consider these three particular inferences for example :

(1) Subject *S* infers from *Joe is a bachelor* to *Joe is unmarried*

(2) Subject *S* infers from *Andrew is a bachelor* to *Andrew is unmarried*

(3) Subject *S* infers from *Roger is married* to *Roger is not a bachelor*

Do these inferences all draw upon the information that *bachelors are unmarried*? Or is it that (1) and (2) draw on the information that bachelors are unmarried, but that (3) does not, and that (3) draws on the distinct information that *married people cannot be bachelors*? Moreover, is it really an *a priori* truth that (1) and (2) draws on the same information? Can it not be the case that what is causally relevant in (1) is the state containing the information that *if Joe is a bachelor then Joe is unmarried*? How information-bearing states are to be individuated, and which are the states causally relevant in an inference, these are questions that can only be answered by careful empirical inquiry. In the absence of such answers it is doubtful whether we can draw any *a priori* conclusions about the syntactic structure of thoughts.

Why should one have thought otherwise in the first place? One explanation is the conflation between the strong and weak notion of causal systematicity. But I think another reason might be that it is all too easy to conflate another distinction : that of "inference" as an argument or proof on the one hand, and as the process of reasoning on the other.[10] We might think of reasoning as a mental process that leads to a rational change of one's beliefs or intentions. But an argument or proof is not a mental process. It is an abstract object that consists of premises and a conclusion, and whose validity does not depend on whether anyone has thought of it. It is of course a psychological fact that we are disposed to find certain patterns of argument as valid. So for example (having mastered the concept of *bachelor*) we are disposed to judge the following arguments as valid : "Gabriel is a bachelor $\therefore$ Gabriel is unmarried", and "Benjamin is a bachelor $\therefore$ Benjamin is unmarried", etc.. It is quite plausible that there is some common mental state $M$ that explains why in both cases we judge the argument as valid, a mental state that also explains why we are disposed to judge arguments of the form "$x$ is a bachelor $\therefore$ $x$ is

[10] For further discussion of this distinction, and the view that rules of argument have very little to do with rules of reasoning, see Gilbert Harman (1986) *Change in View* Cambridge : MIT Press.

unmarried" as valid. However, if one does not distinguish between inference as argument and inference as reasoning, one might take this to show that we have a disposition to *reason* from "$x$ is a bachelor" to "$x$ is unmarried", and that state $M$ is causally efficacious in all instances of such reasoning. The conflation of strong and weak causal systematicity leads straightforwardly to the conclusion that all bachelor thoughts have a syntactic property in common.

Of course, there remains the question of explaining our disposition to find arguments of a certain pattern as valid. One might wonder, how can we explain why we have such a disposition without postulating a language of thought? But the considerations that support such a view are empirical ones such as those based on systematicity and productivity. I have argued that however convincing *these* arguments are, reflections on the nature of concept possession provides no *additional, a priori* argument for LOT, even in the minimal version that Davies has espoused.

# *Chapter Three*

# Three Motivations for Narrow Content

In everyday life, we typically explain what people do by attributing mental states such as beliefs and desires. They belong to a class of mental states that are *intentional*, mental states that have content. Hoping that Johnny will win, and believing that Johnny will win are of course rather different mental states that can lead to very different behaviour. But they are similar in that they both have the same content : what is being hoped for and believed is the very same thing. According to the thesis of externalism that has been defended most notably by Hilary Putnam and Tyler Burge,[1] not all of the contents of our mental states are determined by our intrinsic properties. Instead, the contents of our beliefs and desires are often determined in part by our relations to the environment. They are, so to speak, "wide" contents that are "not in our heads." Although externalism is accepted by most philosophers, many have argued that mental states with wide contents must also have a kind of content wholly determined by the intrinsic properties of the individuals who are in those states. This kind of content is called "narrow content". The aim of this paper is to distinguish between three rather different motivations for postulating narrow content. I argue that, given a certain conception of narrow content that I shall explain below, none of these three motivations justify the postulation of narrow content.

---

[1] See Hilary Putnam (1975) "The Meaning of 'Meaning'" in *Mind, Language, and Reality* Cambridge : Cambridge University Press, pp. 215-271; Tyler Burge (1979) "Individualism and the Mental" *Midwest Studies in Philosophy* Vol. IV, Minneapolis : University of Minnesota Press, pp. 73-121.

## 1. Three Motivations for Narrow Content

Arguments for externalism often rely on familiar thought experiments such as the following one. Jane is an ordinary earthling who is acquainted with water in the normal way, but like many people, she is ignorant of the chemical nature of water, such as the fact that water is made up of $H_2O$ molecules. Nonetheless she has many beliefs about water. For example, she believes that water quenches thirst, and that water puts out fires. But now suppose she has a duplicate twin-Jane who has the same intrinsic properties. Twin-Jane grows up on planet twin-earth where everything is exactly the same as on earth, except that there is no water there.[2] Instead twin-earth has twin-water, a substance that has the same appearance as water. Twin-water tastes just like water, and it quenches thirst in just the same way. But it has a totally different chemical nature, being made of the compound XYZ instead. We are supposed to have the intuition that twin-Jane *lacks* beliefs about water, and that *de dicto* belief ascriptions that use the word "water" and that are true of Jane will not be true of twin-Jane. So unlike Jane, twin-Jane lacks the belief that water quenches thirst, and she lacks the belief that water puts out fire, etc.. Instead, what she believes is that twin-water quenches thirst and puts out fire. Thus Jane and twin-Jane have beliefs that differ in truth-conditions. It is supposed to follow that their beliefs differ in content, despite the fact that they have the same intrinsic properties. If an individual *I* has a mental state *m* with a certain content, and it is metaphysically possible for a duplicate of *I* to lack a mental state with that same content, then I shall say that *m* has a "wide" content.

---

[2] Following everyone else in the literature I shall ignore the fact that twin-Jane and Jane cannot be exactly alike in their intrinsic properties, and assume that this difference is irrelevant to the discussion. The same point that is made with water can be made with other substances which are not present in their bodies.

I shall not challenge the externalist conclusion that many of our mental states have wide contents. But notice that externalism does not imply that *all* intentional mental states have wide contents. Consider my belief that something exists, and my belief that everything is identical to itself. Although some of my possible duplicates might live in communities that speak a somewhat different language, where for example the word "exists" has a slightly different meaning, the claim that my duplicates could lack those two beliefs has very little intuitive support. So if what we mean by there being narrow content is that there are some mental states whose contents are not wide, then I think there is indeed a *prima facie* case for the existence of narrow content.

There is, however, little discussion of such examples by either the friends or foes of narrow content. Those who defend narrow content generally do not argue that there are particular belief predicates of the form "believes that *p*" which are true of an individual and all his or her possible duplicates. Instead, their position seems to be that *every* belief with wide content also has an additional kind of content that is wholly determined by the intrinsic properties of their subjects. So proponents of narrow content agree for example, that the content of Jane's belief that water quenches thirst is wide. Nonetheless they insist that *this very belief* also has a different kind of content that is wholly determined by her intrinsic properties. This is the narrow content of Jane's belief, and although twin-Jane does not believe that water quenches thirst, her belief that twin-water quenches thirst is supposed to have the same narrow content as Jane's belief that water quenches thirst.

In this paper I shall focus on a particular class of intentional mental states - states that have truth-conditions, and in particular beliefs and thoughts. Among the proponents of narrow content, I think it is a common assumption that all beliefs and thoughts have both a narrow and a wide content. For example, on Fodor's account, the truth-condition of a belief or thought is supposed to be a function of its narrow content and some relevant

context. Brian Loar has also argued that beliefs and other intentional mental states have both a social content and a psychological content, only the latter of which is determined entirely by the subject's intrinsic properties.[3] So, whatever narrow content is, I think it is fair to say that it has to satisfy at least these two principles :

(A) *Content is dualistic* : every belief or thought (token) that has a wide content also has a narrow content.

(B) *Narrow content is narrow* : if an individual has a belief (or thought) with a narrow content *N*, then every possible duplicate of that individual also a belief (or thought) with the same narrow content *N*.

These two principles obviously do not explain what it is for a belief or thought to have a narrow content. But they might be regarded as necessary conditions that a satisfactory notion of narrow content have to satisfy. These two conditions are however rather minimal. For example, I might stipulate that *that snow is white* is the narrow content of *all* intentional mental states. Such a notion of narrow content would indeed satisfy principles (A) and (B) : every belief or thought with a wide content also has a narrow content, and the mental states of duplicates do not differ in narrow content. But of course this notion of narrow content is quite useless since it does not explain *anything*. So if the thesis that narrow content exists is to have any interest at all, one would have to show not only that there is a coherent notion of narrow content that satisfies principles (A) and (B). It would also have to be shown that this notion of narrow content has some theoretical interest. In what follows, I shall discuss three different motivations that have led

---

[3] For Fodor's account, see Chapter 2 of Fodor (1987) *Psychosemantics* Cambridge : MIT Press. Loar's argument for narrow psychological content appears in "Social Content and Psychological Content" in Robert Grimm and Daniel Merrill (eds.) (1988) *Contents of Thought* Tucson : Univeristy of Arizona Press, pp. 99-110. David Chalmers has also defended the view that every belief has both a narrow and a wide content. (Chalmers (1994) "The Components of Content" *Manuscript.*)

philosophers to introduce narrow content. The first is that we need to introduce narrow content to account for our first-person knowledge of our intentional mental states. The second is that narrow content is to be understood as the internal component of a mental state with wide content, that which is left of the mental state as we abstract away from its relations to the environment. The third is that narrow content is needed for the formulation of psychological laws. I think it is important to distinguish between these three different motivations, for I think we have as yet no reason to think that there is a single property that performs all three explanatory tasks. Furthermore, I shall argue for a more basic conclusion : the three motivations by themselves provide no reason for thinking that there has to be a theoretically important notion of narrow content that satisfies both principles (A) and (B).

## 2. Narrow Content and Self-Knowledge

Recently there has been some discussion of externalism in connection with first-person knowledge of our intentional mental states. By such self-knowledge, I mean knowledge of our intentional mental states that do not rely on empirical evidence or observation of behaviour. On the face of it, there does seem to be a *prima facie* difficulty in reconciling externalism with our self-knowledge. If the contents of our thoughts are determined in part by our relations to the environment, then one might think that in order to know what we think, we have to find out what our environment is like. But self-knowledge is precisely knowledge that does not come about by empirical investigations. So it seems that we have a dilemma, that either contrary to appearance we do not really know the contents of our own thoughts, or if we do, then externalism is false.

However, the problematic conclusion follows only if we accept the following principle : to know that *p*, one has to know that the conditions necessary for the truth of *p* do indeed obtain. If our purpose is to reconcile externalism with the possibility of self-knowledge, then it seems that narrow content is not really relevant whether we accept this principle or not. First of all, if we reject this principle, there is then no reason why externalism requires that, to know the contents of our own thoughts, we have to know that the environmental conditions necessary for such thoughts do in fact obtain. On the other hand, if the principle is accepted, it threatens to make not just self-knowledge, but empirical knowledge in general, impossible. For example, it implies that Jane cannot know that water puts out fire, since Jane does not know that a liquid made up of $H_2O$ molecules can put out fire, and this has to be true if water does put out fire. But if it turns out that general skepticism is the reason for thinking that externalism cannot be reconciled with self-knowledge, then it is hard to see how postulating narrow content can be of any help. If the skeptic thinks that we cannot know the wide contents of our thoughts because we do not know what our environment is like, then surely this skeptical challenge cannot be met by saying that thoughts have narrow content! So either way, there does not seem to be any reason why we need to appeal to narrow content in reconciling externalism with self-knowledge.[4]

Brian Loar has argued that there *is* some related phenomena that calls for narrow content. According to Loar, "subjective intentionality" describes the fact that we have introspective knowledge of the intentional properties of our thoughts, and that this is something we need narrow content to account for.[5] To use Loar's example, when I entertain the thought that Freud lived in Vienna, there are true judgements that I can make with regard to its intentional properties, e.g. it is a thought about Freud, and it is

---

[4] I am indebted to Daniel Stoljar for a discussion on this topic.

[5] See Loar *op. cit.*, and also Loar (1987) "Subjective Intentionality", in *Philosophical Topics*, Volume XV, No.1, pp. 89-124.

true if and only if Freud lived in Vienna. It is plausible that such judgements constitute knowledge of the intentional properties of my occurrent thought, and that I come to have such knowledge not by carrying out empirical inquiry of the ordinary kind. According to Loar, we do not ordinarily conceive of such intentional properties as extrinsic properties : "From a pre-critical perspective, knowledge of the references of my own thoughts is privileged in a certain way, and that perspective involves no apparent conceptions of external reference relations."[6] Apparently, the claim is that when I judge that, say, my occurrent thought is about Freud, I do not conceive of the property of *being a thought about Freud* as an extrinsic property. But why should this give us any reason to think that narrow content exists? Perhaps the argument is that if I do not conceive of this intentional property as extrinsic, then it has to be the case that the property is an intrinsic one.

But if this is Loar's argument for narrow content, then it is not a very convincing one. First of all, as Stalnaker points out in his discussion of Loar,[7] it is natural to conceive of intentional properties as extrinsic properties even from a commonsensical perspective. Although the externalist thesis is a substantial one, it is one that seems to gather ready conviction even from a pre-theoretic perspective. More importantly, even if there is an individual who judges that his occurrent thought has a certain intentional property, but who does not conceive of the property as extrinsic, it still does not follow that the intentional property so ascribed is intrinsic. One can ascribe a property without knowing the full nature of the property, and it would be just as fallacious to argue that the instantiation of the property of *being water* is independent of chemical constitution, on the ground that one can ascribe the property without conceiving of it as involving molecules.

---

[6] Loar, *op. cit.* page 96.

[7] See Robert Stalnaker (1990) "Narrow Content" in Anderson and Owen (eds.) *Propositional Attitudes : The Role of Content in Logic, Language, and Mind* Stanford : CSLI.

On the other hand, perhaps Loar has something else in mind, when he claims that we do not have an extrinsic conception of the intentional properties we ascribe to our occurrent thoughts. One thing he *might* mean is this : my judgement that my thought is about Freud is not a judgement that is *based on* my beliefs about my relation to the environment. Thus I do not arrive at the judgement by reflecting on the nature of *aboutness*, and inferring from my beliefs about my causal relations to Freud that the thought is indeed about him. This is perhaps why Loar says that "the referential judgement is from a first-person perspective independent of thoughts about causal reference relations." The admittedly correct observation is that I can form direct judgements about the intentional properties of my occurrent thought, without having to consciously infer such judgements from my beliefs about my relations to the environment. I think Loar's position is that there has to be some intrinsic properties about me that explain my judgements of the intentional properties of my occurrent thought, properties which we might take to be the narrow content of the thought. This seems to be the line of thought behind his remarks that the "(objectively non-intentional) properties of object-level thoughts which contribute to explaining why upon reflection they reveal themselves as 'about this and about that' can be counted as the basis of subjective intentionality ... [which is] the disposition of thoughts to reveal themselves 'as intentional' upon reflection."[8]

I think there is a real and interesting issue as to what these intrinsic properties might be. But note that if this is what narrow content comes down to, then such a notion of narrow content will most likely fail to satisfy principle (A). That is, not all beliefs and thoughts with wide contents will have the kind of narrow content that Loar has envisaged. One reason is that subjective intentionality is after all a sophisticated psychological

---

[8] Loar *op. cit.* page 102.

phenomena, present only in creatures capable of reflexive judgements about their own thoughts. Furthermore, even in our own case it is a phenomena that is restricted to only conscious intentional mental states. Certainly an unconscious thought does not normally have the disposition to cause us to judge of *it* that it has some intentional property *P*! So if according to Loar narrow contents are those intrinsic properties that are the basis of subjective intentionality, then it seems quite likely that narrow content is possessed only by those conscious intentional states of creatures capable of reflexive judgements. The mass of our unconscious intentional states, as well as the intentional states of simpler thinking creatures that lack the ability for reflexive thinking, will most likely not be states that have narrow content in this sense. However, I think many philosophers (and perhaps Loar included!) will argue that there is still a need to ascribe narrow content to these mental states for the other two purposes. They would still argue, for example, that psychological laws should also apply to unconscious mental states with wide content, and we need to ascribe narrow contents to such states for the formulation of such laws. To distinguish between these different motivations for narrow content, let us use "subjective content" to designate those properties of our conscious thoughts that enable us to form correct judgements of their intentional properties. I have argued that whatever subjective content is, it is not something that satisfies principle (A).

In fact, if introspection is anything to rely on, there is some reason to think that the subjective content of a conscious thought is some kind of phenomenal property. Conscious and deliberate thinking often *seem* to be a matter of silent speech, and we often *seem* to hear words and phrases in our heads as we think. What this suggests is that there is something which it is like to have a conscious thought, and a difference in the contents of our conscious thoughts can correspond to a difference in what it is like to have those thoughts. There is some plausibility to the idea that such phenomenal properties enable us to identify the intentional properties of our conscious thoughts from a first person

perspective. Whatever the nature of such phenomenal properties might be, at least it seems clear that unconscious occurrent thoughts do not have phenomenal properties. If this is right, it provides some further evidence that subjective content is not a property possessed by all mental states with wide content.

### 3. Narrow Content by Subtraction

So we still have not found an interesting notion of narrow content that satisfies both principles (A) and (B). But why should we think there has got to be such a notion? I think one motivation comes from the powerful and plausible intuition that there has *got* to be a sense in which our mental states are dependent on what is happening *within* us. Although externalism tells us that what we believe and desire are dependent on our relations to the environment, it surely cannot be the case that what we believe and desire are *independent* of our intrinsic nature. Believing that water quenches thirst is clearly not like being three miles away from a burning barn, where the location of the object is all that matters. We cannot just dump an object on twin-earth, and thereby brings it to believe that twin-water quenches thirst. How can we deny that the internal constitution of an object plays *some* role in determining its beliefs and desires? But then it is reasonable to think of having a mental state with wide content as being determined by two factors : an internal factor that depends only on our intrinsic properties, and an external factor that has to do with our relations to the environment. We might then stipulate narrow content as simply the internal component that contributes to being in that mental state with wide content.

Here we might draw an analogy with the case of weight. Having some particular weight is of course an extrinsic property. Nonetheless an object having the weight it has is

a function of its mass and the local gravitational field. What the proponent of narrow content aims to do, is to find some similar way of factoring out the internal component of having a mental state with a wide content. If every mental state with wide content can be thus resolved into an internal and external factor, then this notion of narrow content would indeed satisfy principle (A). Something like this seems to be what Ned Block has in mind when he introduces narrow content as follows :[9]

> Where we have a relation, in certain types of cases we have individualistic properties of the related entities that could be said to ground the relation. If *x* hits *y*, *y* has some sort of consequent change in a bodily surface, perhaps a flattened nose, and *x* has the property of say, moving his fist forward. ... There is a nonrelational aspect of propositional attitude content, the aspect "inside the head," that corresponds to content in the way that moving the fist corresponds to hitting. This nonrelational aspect of content is what I am calling narrow content.

The idea is that whether the relation "*x* has a belief with wide content *y*" obtains between an individual and a certain content does depend on the extrinsic properties of the individual. But one might try to factor out those aspects of the relation that is "inside the head".

However, many philosophers have pointed out that such a move is not entirely innocent. Hilary Putnam has gone even further to argue that this project of factoring a belief is doomed to fail. Putnam argues that "there is no one physical state or one computational state that one has to be in to believe that there is a cat on the mat."[10] He considers and rejects various candidates (e.g. perceptual prototypes, conceptual roles) for

---

[9] Ned Block (1986) "Advertisement for a Semantics for Psychology" in *Midwest Studies in Philosophy* Vol. X, Minneapolis : University of Minnesota Press, pp. 615-678.

[10] Putnam (1992) "Why Functionalism Didn't Work" in John Earman (ed.) *Inference, Explanation and Other Philosophical Frustrations* Berkeley : University of California Press. See also Putnam (1988) *Representation and Reality* Cambridge : MIT Press.

the narrow content of this belief, and argues that none is to be found. I think that Putnam is indeed correct on this point, but it is not clear why a proponent of narrow content should be troubled by this line of argument at all. For this criticism presupposes that *tokens* of beliefs with the same wide content must all have the same narrow content. However, there is no reason why a proponent of narrow content is committed to this assumption. Consider again the analogy with weight : having a weight of five grams, like having the belief that water quenches thirst, is an extrinsic property. But is there a single intrinsic property that all objects must have in order to weigh five grams? Objects that weigh five grams can have different masses and differ in other intrinsic properties in all sorts of ways. Of course, they do share the intrinsic property of having a non-zero mass. But having a non-zero mass does not contribute to explaining why an object has the weight it does, even given the strength of the gravitational field it is in. It is the property of having the particular mass it has that is the explanatory internal property. This is also true of Block's example of hitting, where the internal aspects of different instances of hittings can be different, even if the subjects and the patients are the same.

Similarly, a proponent of narrow content can cheerfully agree that there is no single interesting intrinsic property shared by all those who believe that there is a cat on the mat. It is only natural that people who share this belief differ in all sorts of ways in their computational states and internal constitution. They do share the property of having *a* belief of course, and perhaps it might be argued that this is an intrinsic property. After all, although externalist arguments show that for many contents, having a belief with those contents depends on how one is related to the environment, they certainly do not show that having a belief with *some* content is also an extrinsic property. But even if having a belief is an intrinsic property, this is not what narrow content is. The narrow content of a particular belief token is determined by the internal factor that goes toward explaining why the subject has that particular belief. Two belief tokens with the same

wide content might nonetheless have different narrow content. But this is no more problematic than there being two things having the same weight but different mass.

I can think of no strong objection to this way of understanding narrow content, where narrow content is primarily a property of mental state tokens and not of state types. But I think the following point is worth noting. Recall that narrow content is supposed to satisfy two principles :

(A) *Content is dualistic* : every belief or thought (token) that has a wide content also has a narrow content.

(B) *Narrow content is narrow* : if an individual has a belief (or thought) with a narrow content *N*, then every possible duplicate of that individual also a belief (or thought) with the same narrow content *N*.

I have argued that subjective content is not quite narrow content because it does not satisfy principle (A). On the current proposal where each belief token with wide content is to be factored into an internal and external component, principle (A) is indeed satisfied if narrow content is identified with the internal component. But we do not know whether this notion of narrow content satisfies principle (B) or not. We do know for example that Jane's belief that water quenches thirst is determined by a combination of internal and external factors, and the narrow content of that belief token is whatever intrinsic property it is that in part explains why Jane has that particular belief. Call this property *P*. Whatever *P* is, by definition it is an intrinsic property that is shared by twin-Jane. But it does not follow that *P* also has to be the narrow component of twin-Jane's belief that twin-water quenches thirst. It is true that twin-Jane's belief is also a function of two factors. But we have as yet no reason to think that the internal factor of this belief also has to be the property *P*. Or for that matter, we have no reason to think that *P* has to be

the narrow component of *any* belief-token of twin-Jane. If the narrow content of Jane's belief is to be identified with the property *P* or is something that is determined by *P*, the present account of narrow content leaves it open that none of the belief tokens of twin-Jane has the same narrow content as Jane's belief that water quenches thirst. So narrow content might turn out not to be narrow!

Notice that the problem is not due to the fact that we have not specified what *kind* of property *P* is supposed to be. For example, if a conceptual role theory is correct, then we know that property *P* has to do with the internal conceptual role of Jane's belief. But still it does not follow that the conceptual role that goes toward explaining why Jane believes that water quenches thirst is the same conceptual role that goes toward explaining why twin-Jane believes that twin-water quenches thirst.

Perhaps an analogy might help. Imagine Jane whistling and walking about the room with Matthew nearby. Matthew finds the whistling extremely annoying but he does not mind her pacing to and fro. On twin-earth, twin-Jane has the same intrinsic properties and also whistles and walks about the same way. But twin-Matthew finds her pacing unbearable instead even though he has no trouble with her whistling. Now Jane has the extrinsic property of irritating Matthew which twin-Jane lacks, and twin-Jane has the property of irritating twin-Matthew that Jane lacks. Obviously, the internal component of Jane's irritation of Matthew is different from that of twin-Jane's irritation of twin-Matthew.

Perhaps one might ask, why is it important that these two belief tokens should have the same narrow content? In a way, this is not really important, if all that one aims to do is to identify the internal and external factors that having the belief tokens consist in. The success of such a project does not require that narrow content have to satisfy

principle (B). However, one of the main motivations for introducing narrow content in the first place is that it is needed for psychological explanation. On this line of thought, externalism shows that the mental states of duplicates need not share the same wide content. It is then argued that this makes the wide contents of mental states unsuitable for psychological explanations, and that it is their narrow contents that psychological explanations should appeal to. On the present proposal, given an individual *I* who is in a mental state *m* with wide content *W*, the narrow content of *m* is *stipulated* as the internal component that explains why *I* is in state *m* with wide content *W*. But as we have seen, it just does not follow from such a stipulation that the mental states of a duplicate of *I* will have the same narrow content. Of course, I have no argument that they will be different either. But I think it is common ground that if narrow content is relevant to psychological explanation, then it has to satisfy principle (B). So, without a more detailed account of how we are to factor a mental state into an internal and external component, it is not clear whether this present notion of narrow content is of any use in providing psychological explanations.

Furthermore, there is certainly no *a priori* reason to think that such a notion of narrow content has an important role to play in psychology, even if it satisfies principle (B). In the case of weight, its internal component does turn out to be an important physical quantity. But it is worth noting that if we have a scientific theory that makes use of some extrinsic property or relation, even if we can factor out the relevant internal components of instances of such properties and relations, there is in general no *a priori* guarantee that such components are of any interest to that theory.

This is particularly clear in cases where we are dealing with a complex system, where the theoretical concepts of interest are mainly concepts of extrinsic properties or relations of dependence. Consider extrinsic properties in economics such as having a

particular price, or having some particular level of income. There are economic generalizations about such properties and their like, and of course we do not want to say that the intrinsic properties of an object is completely irrelevant to its instantiation of such properties. It is not clear how we might factor out the internal contribution that an object makes to its having the price it has. Nonetheless in the formulation of economic generalizations, we would not expect to appeal to those intrinsic properties thus factored out, or to appeal to the concept of narrow price. Why should we think that it is different in the case of psychology? One might hold that there are generalizations that relate different mental states with wide contents, without thereby committing to the view that there are generalizations to be discovered by factoring such intentional mental states into their internal and external components. So why should we think that psychological explanations have to make use of a notion of narrow content that satisfies principles (A) and (B)? This is the issue to which we now turn.

## 3. Narrow Content and Psychological Explanation

Arguments for the need of narrow content in psychology originally came as reactions to externalist arguments such as those based on the twin-earth thought-experiments. But why do these thought-experiments show that mental states that have wide contents also have narrow contents? What externalism shows is that intentional mental properties, such as having a belief with some particular wide content, are extrinsic properties that do not supervene on a believer's intrinsic properties. But it does not follow from their being extrinsic properties that they are irrelevant to psychological explanation. Consider for example David Lewis's proposal that to causally explain an event is to provide information about its causal history.[11] This seems to be a plausible proposal of

---

[11] David Lewis (1986) "Causal Explanation" in *Philosophical Papers : Volume II* Oxford : Oxford University Press.

what it is to provide a causal explanation. We have a good causal, psychological explanation of an event *e* to the extent that the information provided about the causal history of *e* is psychological, and relevant to the purpose at hand. But then it would seem that extrinsic properties such as having a certain mental state with such and such a wide content can be relevant. I can provide information about the causal history of an event *e* by telling you that the history includes a mental state of attitude type *A* with a wide content *C*, and by receiving this information it allows you to rule out other possible casual histories of *e*. Thus I might causally explain why Jane went out of her office, by saying that she wanted to drink some water from the faucet outside. My explanation would of course be wrong if the causal history of Jane's action does not include a desire for water, but if it is does, then surely this is a case of successful psychological explanation. So, if mental states with wide contents can be relevant to psychological explanations, why do we have to postulate that they have an additional kind of content for explanatory purposes?

Of course, it is not unreasonable to think that when it comes to more detailed *scientific* explanations in psychology, we would also have to appeal to psychological states and processes that are determined only by the subject's intrinsic properties. But the fact that intrinsic properties are relevant to psychological explanations do not show that we need to assign narrow contents to mental states with wide contents. Consider the computer model of the mind, which I assume provides a faithful picture of much of cognitive science.[12] On this conception, a major task of psychology is to identify the computational processes and representations that are causally involved in our mental lives. On the semantic or informational level, we might try to find out what information is made available and processed by those computational mechanisms, what properties of the

---

[12] For discussion of this model, see Jerry Fodor and Zenon Pylyshyn (1988) "Connectionism and cognitive architecture : a critical analysis" *Cognition*, 28, pp. 3-71; Ned Block (1990) "The Computer Model of the Mind" in Daniel Osherson and Edward Smith (eds.) *Thinking : An Invitation to Cognitive Science, Volume 3* Cambridge : MIT Press.

organism and the environment are mentally represented. But we might also study these computational algorithms and representations at a formal level, abstracting away from semantic or informational properties. Finally, we might also try to discover how these representations and algorithms are related to the underlying biological tissues, such as our neural structures. One would have thought that at the formal level, the same computational algorithms and representations are to be found in for example Jane, twin-Jane and their duplicates. It is surely the task of psychology to discover these narrowly individuated computational properties. These properties are therefore good candidates for providing psychological explanations that do not appeal to extrinsic properties. Furthermore, such computational explanations do not proceed by attributing narrow contents to mental states that have wide contents.

Of course, explanations do come at different levels, and one might think that there is a need for psychological theorizing at a level more general than particular computational mechanisms. Instead of considering Jane and twin-Jane, we might consider perhaps cousin-Jane on twin-earth whose mental representations and processes are rather different. Perhaps the sentences of her language of thought, if there is one, have a somewhat different syntax, and that the rules that operate on those representations are different from that of Jane as well. One might argue, surely one should not preclude *a priori* that there *might be* a level of psychological explanation that applies to both Jane and cousin-Jane? But the mental states of Jane and cousin-Jane have different wide contents, and it is supposed to be the case that their computational representations and algorithms are rather different even at the formal level. Presumably we do want a theory of psychology that captures the generalizations that apply to both of them. Would this not be a case for attributing narrow contents to their mental states?

However, I think this line of reasoning takes the idea of a "level" more seriously than necessary. There is no single level that is *the* formal level, since algorithms and representations can be classified by their formal properties at different levels of abstraction and in ways that cross-cut each other. Formally different computational algorithms can for example be classified according to whether they are deterministic or not, and different systems of representations can be classified together according to the formal features of their grammars. One might provide generalizations about the properties of a class of algorithms and representations in terms of such and similar features, and this might prove to be of some importance in psychological theorizing. As it stands, the case of Jane and cousin-Jane is under-described. The fact that there are *some* respects in which their computational architectures are formally different does not preclude there being formal generalizations that apply to them both.

Similarly, although two individuals can differ in the wide contents of their thoughts, I think it will be a mistake to conclude from this fact alone that there cannot be (wide) intentional generalizations that apply to them both. ( So even if there are no interesting formal generalizations that apply to both Jane and cousin-Jane, it does not follow that there are no intentional ones that do. ) However, some such assumption seems to have been made by Fodor's former self in his previous arguments for narrow content. The line of thought is that (a) psychology should provide intentional generalizations that subsume the intentional mental states of both Jane and twin-Jane, but (b) this would not be possible unless their intentional mental states have narrow content. If I understand Fodor correctly, I think he no longer accepts (a), even though he still subscribes to (b). We shall come to Fodor's current view shortly, but for the time being, let us focus on (b) instead. What might be the reason for thinking that the same intentional generalization cannot apply to both Jane and twin-Jane? Some hints are to be found in this revealing passage :[13]

> What the Putnam/Burge examples show is that the broad, folk-theoretic notions of semantic property exhibit a previously unnoticed relativization to context. Narrow content wants to generalize over the contexts to which broad content relativizes, hence permitting psychological laws which hold without respect to context.
>
> Abstracting from context sensitivity is a standard way of achieving scientific generalization. We could have done physics with *weight*, but the price would be context sensitivity in the laws of mechanics. *Mass* generalizes over the contexts to which weight relativizes and is the preferred parameter for precisely that reason. Such precedents would motivate narrow content even if metaphysical arguments for supervenience didn't.

The idea seems to be that if scientific laws subsume states or objects by their extrinsic properties, properties that are "context-sensitive", then such laws will fail to hold across contexts where the extrinsic properties differ. We want the same intentional psychological laws to apply to both Jane and her duplicates, but the worry is that if intentional laws subsume mental states by their wide content, then they might apply to Jane but not twin-Jane because of differences in the wide content of their mental states. If there is to be an intentional law that subsumes both Jane's belief that water quenches thirst and also twin-Jane's belief that twin-water quenches thirst, it has to be the case that these two beliefs have some other content in common that does not depend on the environment.

---

[13] From Fodor's reply to Stalnaker, in Loewer and Rey (1991) *Meaning in Mind : Fodor and His Critics* Oxford : Blackwell, page 318. The same line of argument also appears in his paper "A Modal Argument for Narrow Content" (*The Journal of Philosophy* Volume LXXXVIII, No. 1. pp.5-26.). In that paper, Fodor argues that for psychological purposes, twin's thoughts have the same causal powers. Furthermore, he claims that unless a psychologist attributes narrow content to those thoughts, "his theory misses generalizations, namely, all the generalizations that subsume me and my twin. Good taxonomy is about *not* missing generalizations." Again the assumption is that if twins differ in the wide contents of their mental states, they cannot be subsumed by the same intentional generalization.

But if this really was Fodor's worry, then it seems to be based on a simple confusion. It is one thing to claim that psychological laws subsume mental states by extrinsic properties whose instantiation depends on the context. It is surely quite another to say that the laws formulated in terms of such properties are themselves context-dependent and have limited validity. The second does not follow from the first at all. In fact this should be clear from Fodor's own comparison with classical mechanics. Classical mechanics tells us that if an object with initial velocity $u$ moves with a constant rate of acceleration $a$, then after a period of time $t$ it will have travelled a distance $d$ given by the formula $d = ut + \frac{1}{2}(at^2)$. Like weight, the initial velocity and rate of acceleration of an object are its extrinsic properties, and are thus "context-sensitive" in Fodor's sense. Yet this particular law of motion specified in terms of such properties is surely not context-dependent in any interesting sense. It holds with respect to contexts where objects may vary in their initial velocities and rates of acceleration in all sorts of ways. There is no fear of loss of generalization because it subsumes objects by their extrinsic properties, nor should we say that in additional to an object's initial velocity and rate of acceleration it also has a narrow velocity and a narrow acceleration!

The same point can be made with regard to Fodor's worry that the same intentional laws will fail to apply to duplicates if they subsume mental states by their wide content. As many authors have argued, intentional generalizations are generalizations that quantify over contents. This includes Fodor himself who repeatedly points out that the formulation of such generalizations do not mention particular contents at all. In *Psychosemantics*, he writes,[14]

---

[14] Fodor (1987) *Psychosemantics* Cambridge : MIT Press, page 70. Paul Churchland makes the same point in Churchland (1981) "Eliminative Materialism and the Propositional Attitudes" *Journal of Philosophy* 78, No. 2, pp. 67-90.

> [P]sychological theories typically achieve generality by *quantifying over* the objects of the attitudes. In consequence, many of the most powerful psychological generalizations don't care about content per se; what they care about is only relations of *identity and difference* of content.

But if intentional generalizations do not "care" about particular contents, why can't we have the same intentional generalization subsuming the different intentional states of the twins? One such generalization that Fodor mentions in the first chapter of *Psychosemantics* is this : "if someone believes that $Fa$, then *ceteris paribus*, that person believes $\exists x(Fx)$." Let us assume that this is indeed a true generalization. One would have thought that this will be a generalization that applies to both Jane and twin-Jane. Jane believes that water quenches thirst; twin-Jane believes that twin-water quenches thirst. Presumably the generalization predicts that they both believe that something quenches thirst. If as Fodor says, psychological generalizations do not mention particular contents but quantify over them, there is then no reason to believe that such generalizations cannot subsume the different intentional states of Jane and twin-Jane, and maybe that of cousin-Jane also.

I should perhaps explain why I have chosen Fodor as my target here even though he has recently come to the conclusion that perhaps psychology can make do without narrow content. The reason is that despite his change of position, he still seems to hold that twin earth cases show that broad intentional laws will miss generalizations. In his Jean Nicod Lectures, he writes, "Twin cases say : if you insist that computationally implemented intentional laws be broad, you will miss generalizations in virtue of which my psychology is the same as that of my computationally identical twin."[15] But why should this not be a problem for psychology? If I understand Fodor's response correctly, his answer is that it is alright even if psychology misses such generalizations. This is because

---

[15] Fodor (1993) *The 1993 Jean Nicod Lectures* Manuscript.

(i) psychology is a special science whose intentional laws are *ceteris paribus* laws, and (ii) twin earth contexts are ones where the *ceteris paribus* clauses of intentional laws fail, and (iii) a special science need not provide generalizations that deal with contexts in which *ceteris paribus* clauses fail.[16] However, if the point I have just made is correct, there is no reason to think that twin-earth cases pose a problem for the generality of intentional laws in the first place.[17] Although Jane and twin-Jane differs in the wide contents of their mental states, there *is* a straightforward sense in which they have the same psychology : their intentional mental states are subsumed by the same set of intentional laws. Not only that, as suggested earlier the computational states and processes that we find in both of them, as characterized in formal terms, are exactly the same. One does not therefore have to postulate narrow content then, to satisfy the desire for psychological explanations that appeal to only intrinsic properties, or for intentional laws that subsume the mental states of individuals on both earth and twin-earth.

## 4. Conclusion

The main point I have been arguing in this paper is that there is no straightforward inference from externalism to narrow content. I do not want to deny however, that perhaps some of our mental states do have contents that are determined by

---

[16] In Fodor *op. cit.* he writes, "Intentional psychology is a special science. So its laws are ceteris paribus laws. And *ceteris paribus laws tolerate exceptions* as long as the exceptions are unsystematic. ... Occasional and fortuitous Twins might be tolerable as consequences of failures of ceteris paribus clauses to be satisfied. I claim that, if you think about what Twins are required to be like, and you think about what the world actually is like, you'll see that if there are Twins, that's morally certain to be accidental."

[17] Could it be that Fodor has changed his mind about the nature of intentional laws? That perhaps they do not quantify over contents after all? But this seems unlikely because first, one would have thought that he would *say* so. Second, such a position would have it that there are generalizations such as, say, "For all $X$ if $X$ believes that snow is white, then $X$ believes that something is white." But we can replace "snow" by any other mass term and get indefinitely many true generalizations. Clearly this cannot be an accident. It is hard to see how one can deny that these generalizations are not instances of still more general ones.

our intrinsic properties. As suggested in the beginning of this paper, it might be that externalism is not true of some of our beliefs. Furthermore, I have not addressed the issue concerning externalism with regard to the contents of perceptual experiences.[18] Finally, I think it is not unreasonable to think that some of our computational representations also have contents that are determined by our intrinsic properties. Perhaps there are subpersonal states that can be said to represent the distribution of light as detected by our retinal cells. It is also reasonable to think that in any complicated computational system such as ours, there are states which monitor other states of the internal environment, such as representing whether a certain module has successfully carried out some operation. If it makes sense to ascribe content to such states at all, then these would be good candidates of states whose contents do supervene on their subject's intrinsic properties. What I have been trying to argue against, however, is the view that there has to be a useful notion of narrow content that satisfies principles (A) and (B). I think we have good reasons for thinking that subjective content does not satisfy principle (A), and I pointed out that narrow content introduced as the internal component of a wide mental state need not satisfy principle (B). Finally, considerations based on psychological explanations do not motivate such a notion of narrow content either. Of course, I have not *shown* that there cannot be a theoretically important notion of narrow content that does satisfy these two principles. But I think this is as it should be. Narrow content is a theoretical concept, and there is no better argument for the need of such a concept other than by embedding it within a theory and show how it bears fruit. What I have been trying to resist, are relatively *a priori* arguments intending to show that the project has to proceed in some particular way.

---

[18] For a discussion, see for example Burge, *op. cit.*, and Davies (1991) "Individualism and Perceptual Content" in *Mind* October issue, and also the reply from Segal in the same issue.