

Volume 5 Number 2  
2009

ISSN 1807-9792

# abstracta

*Linguagem, Mente & Ação*

**Making "Reasons" Explicit:  
How Normative is Brandom's Inferentialism?**  
Daniel Laurier

**On The Interpretation Of Hume's Epistemology**  
J. P. Monteiro

**Editorial**  
**Second European Graduate School:  
Philosophy of Language, Mind and Science**  
Albert Newen, Raphael van Riel & Michael Sollberger

**Semantic Reference Not By Convention?**  
Jessica Pepp

**Prospects for an Intentionalist Theory Of Self-Deception**  
Kevin Lynch

**Synaesthesia and The Relevance of  
Phenomenal Structures in Perception**  
Michael Sollberger

**ABSTRACTA**  
*Linguagem, Mente e Ação*

ISSN 1807-9792

Volume 5 Number 2  
2009

**Editors**

André Abath  
Leonardo de Mello Ribeiro  
Carlos de Sousa

**Guest Editors**

Albert Newen  
Raphael van Riel  
Michael Sollberger

**Executive Editors**

Jules Holroyd  
Gottfried Vosgerau

**Associate Editors**

Abílio Azambuja  
Miquel Capo  
José Edgar González  
Vanessa Morlock  
Olivier Putois  
Erik Rietveld  
Giuliano Torrenço

## TABLE OF CONTENTS

<b>Making "Reasons" Explicit: How Normative is Brandom's Inferentialism?</b>	<b>79</b>
Daniel Laurier ( <i>Université de Montréal</i> )	
<b>On The Interpretation of Hume's Epistemology</b>	<b>100</b>
J. P. Monteiro ( <i>Universidade de São Paulo</i> )	
<b>Editorial – Second European Graduate School: Philosophy of Language, Mind and Science</b>	<b>113</b>
Albert Newen ( <i>Ruhr-Universität Bochum</i> ), Raphael van Riel ( <i>Ruhr-Universität Bochum</i> ) & Michael Sollberger ( <i>Université de Lausanne</i> )	
<b>Semantic Reference Not By Convention?</b>	<b>116</b>
Jessica Pepp ( <i>University of California, Los Angeles</i> )	
<b>Prospects for an Intentionalist Theory of Self-Deception</b>	<b>126</b>
Kevin Lynch ( <i>Warwick University</i> )	
<b>Synaesthesia and the Relevance of Phenomenal Structures in Perception</b>	<b>139</b>
Michael Sollberger ( <i>Université de Lausanne</i> )	

**MAKING "REASONS" EXPLICIT:  
HOW NORMATIVE IS BRANDOM'S INFERENTIALISM?**

**Daniel Laurier**

**Abstract**

This paper asks whether Brandom (1994) has provided a sufficiently clear account of the basic normative concepts of commitment and entitlement, on which his normative inferentialism seems to rest, and of how they contribute to explain the inferential articulation of conceptual contents. I show that Brandom's claim that these concepts are analogous to the concepts of obligation and permission cannot be right, and argue that the normative character of the concept of commitment is dubious. This leads me to replace Brandom's conception of inferential relations as relations between deontic statuses with one according to which they are to be seen as relations between entitlements and acknowledgements of commitments.

*1. Introductory Remarks*

Brandom (2001) draws an interesting analogy between the status of modal concepts and that of normative concepts. He points out that while early (roughly, pre-Kripkean) naturalists questioned the intelligibility of modal concepts and tried either to dispense with them or to explain them in non-modal terms, it is striking that contemporary (roughly, post-Quinean) naturalists no longer see modal concepts as problematic and freely resort to them in their various explanatory projects, including the project of giving an account of meaning and intentionality. This is a welcome development of course, since according to a Kantian-Sellarsian argument on which Brandom puts much emphasis, the availability of non-modal concepts presupposes the intelligibility of modal concepts. Contemporary naturalists, however (as Brandom's story continues), are still highly suspicious of relying on the use of normative concepts in giving an account of intentionality (or indeed, of anything). But their reluctance is unwarranted, since (according to Brandom) the very same Kantian-Sellarsian argument which shows that the availability of non-modal concepts presupposes the intelligibility of modal concepts can be adapted to show that the availability of non-normative concepts presupposes the intelligibility of normative concepts. Hence, we should feel free to use normative concepts in our accounts of meaning and intentionality.

But just as admitting the legitimacy and irreducibility of modal concepts doesn't eliminate the need for an account of how they work and what they mean, admitting the legitimacy and irreducibility of normative concepts doesn't eliminate the need for a corresponding account of how they work and what they mean. Much progress has been made, in the last 50 years (or so), in our understanding of modality, but I think it is fair to say that no progress of comparable magnitude has been made in our understanding of normativity. This is basically how Rosen (2001) responded to Brandom's article: by asking whether our grasp of normativity is any better than our grasp of meaning and intentionality, and if so, whether it is sufficiently better for it to be used in giving an account of the latter.

In the present paper, I am asking the more specific question whether Brandom (1994) himself has provided a sufficiently clear account (i) of those basic normative concepts on which the normative inferentialism he has developed seems to rest, and (ii) of how they contribute to explain the inferential articulation of contents.<sup>1</sup> I will be raising a number of worries and misgivings which, taken together, will make it hard to avoid the modest conclusion that we do not have a firm grasp of the relevant normative concepts and do not understand what the resulting account of intentionality is supposed to be.

## *2. Brandom's Basic Normative Concepts*

Brandom's leading idea is that conceptual thought is inseparable from discursive practice, where this is conceived, most basically, as a kind of implicitly normative practice in which the fundamental moves (i) confer certain deontic statuses on the participants and (ii) are inferentially articulated, in the sense that they both count as the giving of reasons and are themselves in need of reasons, and thereby qualify as assertions. He recognizes two basic kinds of deontic statuses: commitments and entitlements. These are singled out as the two primitive normative concepts in terms of which his account of discursive practice (and thus, of intentionality) is to be framed (1994, pp. 159-166). They are to be seen as normative

---

<sup>1</sup> It is worth stressing that what I am going to say has no bearing on Brandom's much wider program of analytic pragmatism, as expounded in his recent 2008 book, except to the extent that his normative inferentialism has somehow been absorbed into it; and even then, I will merely be questioning Brandom's official normative account of inferential relations, but not the very idea of giving such an account. My aim is clarification, not criticism.

insofar as "[d]oing what one is committed to do is appropriate in one sense, [and] doing what one is entitled to do is appropriate in another" sense (1994, p.159).

But if one were asked to give examples of normative concepts, the concepts which would probably first come to mind are such concepts as "ought" and "reason", and not the concepts of commitment and entitlement. A question thus arises as to what it is that makes them normative and what kind of normative concepts they are; in other words, one would want to be told exactly how the concepts of commitment and entitlement relate to the core normative notions of reason and obligation, and thereby, to be told in what senses it is "appropriate" to do what one is committed or entitled to do. And it is unclear, on reflection, how they are related to these familiar normative concepts.

Brandom declares (1994, p. 160) that "[c]ommitment and entitlement correspond to the traditional deontic primitives of obligation and permission", and it is tempting (and easy) to take them as simple variants of these familiar concepts; which would in turn make it easy to connect them with the concept of reason, since (as it is widely held) what one ought to do is nothing but what one has (most) reason to do. He goes on to suggest (1994, p. 160) that just as being obliged to do something can be defined as not being permitted not to do it (and being permitted to do something as not being obliged not to do it), being committed to do something could be defined as not being entitled to not doing it (and being entitled to do something as not being committed to not doing it), which certainly reinforces the impression that "commitment" and "entitlement" are little more than other words for "obligation" and "permission". However it seems this can't be right, even if it should turn out that commitment and entitlement are interdefinable in the way that obligation and permission are (which also is questionable).

For it is abundantly clear (from the way in which Brandom uses these terms) that commitment doesn't entail entitlement, while no one would want to deny that obligation does entail permission; from which it follows that either commitment doesn't entail obligation or permission doesn't entail entitlement. And it does seem intuitively implausible to suggest that permission to do something entails entitlement to do it, *if* (as seems to be the case) being entitled to do something requires having a reason or being justified to do it, since one clearly may have no reason to do what one is permitted to do. It

doesn't look as implausible to hold that entitlement to do something entails permission to do it, but combining this with the claim that being committed to do something entails not being entitled not to do it would lead to the conclusion that being committed to do something entails not being permitted not to do it (i.e., having the obligation to do it), and hence, would definitely establish that permission doesn't entail entitlement. At this point, it looks as if we could maintain that commitment entails obligation and entitlement entails permission, while denying that commitment entails entitlement, and yet hold that being committed to do something entails not being entitled not to do it. But this last claim must nonetheless be rejected because it would make it impossible for one to have incompatible commitments (e.g., to be simultaneously committed to do something and not to do it) while still being entitled to (discharge) at least one of them; and this is something Brandom explicitly recognizes to be possible. Furthermore, it must be observed that while there is a familiar distinction between *prima facie* and (what some call) *ultima facie* obligations (and permissions), no such distinction seems to apply to commitments (though some such distinction, as will be seen below, seems to apply to entitlements).

So it remains unclear how commitments and entitlements are related to obligations and permissions, and in what senses it is "appropriate" to do what one is committed or entitled to do. This suggests it would be more promising to inquire into how these concepts relate to the concept of "reason", even though it is quite unclear, at first sight, how this could deliver *two different* senses of appropriateness. Although it sounds natural to suppose that doing what one is entitled to do is appropriate because one always has some reason to do what one is entitled to do, it seems implausible to hold that doing what one is committed to do is appropriate because one always has a reason to do what one is committed to do, unless "reason" is here used in a quite different sense. It does seem intuitively false that one always has some reason to do what one is committed to do. Suppose you are committed to kill your neighbor because you said you will do it, or perhaps because it is the only way for you to reach a certain goal. Does that mean you have a reason to do it? If it does, it certainly is not in the sense in which you have reason to do what you are entitled to do, and it is unclear whether and in what sense "reasons" of this kind will count as normative. Moreover, it is puzzling, in this context, to have to turn to the

concept of reason, since (at least in Brandom's view) both reasons and what they are reasons for are *essentially* conceptually articulated in a way that commitments and entitlements in general are not (only *discursive* commitments and entitlements are conceptual, but not all commitments and entitlements are discursive), and we are supposed to be looking for normative concepts that could be used in explaining (*inter alia*) what it is to be conceptually articulated. But let us put these worries aside for a while, and ask how discursive deontic statuses relate to reasons.

A discursive practice is one in which certain performances (e.g., uttering a sentence) count as assertions. To make an assertion, in Brandom's view, is at once to acknowledge and undertake a certain commitment which counts as "doxastic" in virtue of the fact that it is inferentially articulated in the sense that acknowledging such a commitment is both to give a reason and to do something which can be seen as in need of reasons. If I understand him correctly, Brandom's strategy is to explain what it is for such a doxastic commitment to have a certain conceptual (propositional) content in terms of the inferential relations which link it to other doxastic commitments. In other words, inferential articulation is to be seen as pertaining primarily to relations among (discursive) deontic statuses, and only derivatively to relations among conceptual contents such as propositions. Loosely speaking, the idea is that instead of saying, for example, that one cannot be committed to  $p$  without being committed to  $q$  because (in virtue of the fact that)  $p$  entails  $q$ , we are, on the contrary, to be led to see that  $p$  entails  $q$  because (in virtue of the fact that) one cannot be committed to  $p$  without being committed to  $q$ .

On the face of it, the foregoing (admittedly *very* rough) characterization of discursive practice appeals not only to the (putatively) normative concept of commitment, but also to the even more basic normative concept of a reason. It is thus somewhat surprising that *Making It Explicit* doesn't give the latter any definite "official" status (despite making extensive use of it). The purpose of the rest of this paper is to suggest that there is still a lot of substantial work to be done before we could claim to have a firm enough grasp of this concept, and of how it relates to the concepts of (discursive) commitment and entitlement, for them to be relied on in an account of intentionality (and especially, of conceptual contents).

Let us first focus on entitlements. As will have been noticed, the characterization of discursive practice that has been offered doesn't mention them at all. So where are they? The only way in which they (implicitly) get involved in this characterization is through the close relationship they seem to have with reasons (and commitments). To say that doxastic commitments are in need of reasons is to say that it may be asked what it is that entitles one to them, and to answer that question, i.e., to give a reason, is to acknowledge (and undertake) a further doxastic commitment. Thus, at least in the context of discursive practice, entitlements essentially are entitlements *to* discursive (and in the first instance, doxastic) commitments (which makes them some sort of "higher-order" deontic statuses).

I just pointed to the intuitive connection between reasons and entitlements by saying that asking for a reason for a doxastic commitment is asking for what it is that entitles one to this commitment. But this can easily be seen to be ambiguous. There is a sense in which to say that something entitles one to a certain doxastic commitment, amounts to saying that this something makes it the case that one enjoys the deontic status of being entitled to this commitment, period. When we see it in this way, being entitled to a doxastic commitment is tantamount to having ("all things considered") sufficient reason for this commitment. But there is another way to understand the claim that something entitles one to a certain doxastic commitment, according to which it says that the something in question contributes positively to, or counts in favor of, one's being entitled to this commitment, or in other words, that the something in question is what is often called a *pro tanto* reason for this commitment. To read it in this way is to see it as making an irreducibly relational, non-detachable use of the verb "to entitle". What I mean by this, is that one's having a *pro tanto* reason for a certain doxastic commitment doesn't make it the case that one is entitled to this commitment. It would certainly be possible to say that it makes it the case that one is *prima facie*, or *ceteris paribus*, entitled to this commitment; but this would amount to introducing a quite different sort of entitlements (and one which could apparently only be understood in terms of this contrast between a detachable and a non-detachable sense of "to entitle"). I don't mean to suggest that Brandom is unaware of this distinction; on the contrary, there are many indications that he intends his usage of the concept of entitlement to do double duty and cover both cases. The problem remains, however, that it is not

always clear exactly how it is used; and in any case, failing to keep track of this distinction seems to mask the real structure of the view being propounded.

Let us now turn to the concept of commitment, and ask (i) how it relates to the concepts of reason and entitlement, and (ii) whether the way in which it relates to them is apt to reveal its (putatively) normative character.

One seemingly obvious thing is that a commitment is always something for which it makes sense to ask whether one is entitled to it, and thus something for which it makes sense to ask whether there is any (or sufficient) reason. This suggests that one could not be committed to anything without thereby being committed to being entitled to this commitment. It is unclear, however, whether this (even if it turned out to be the case) would be enough to display the normative character of commitments. For the concept of commitment to clearly qualify as normative (deontic), *perhaps* it must be the case not only that if something counts as a commitment then it necessarily is something for which reasons may be asked/given (i.e., something to which one may be entitled), but also that if something is such that reasons may be asked/given for it (i.e. such that one may be entitled to it) then it necessarily counts as a commitment. Now, I do think it is arguable that the concept of something for which (normative) reasons may be asked/given is the concept of something *intentional* (such as an intentional act/attitude), and that there is thus an internal, conceptual link between normativity and intentionality. But this could not be of any help in the present context, since the plan is to explain intentional acts/attitudes in terms of commitments, and not the opposite. Furthermore, even if it is granted that intentional *attitudes* could somehow be seen as discursive commitments, it seems implausible that intentional *acts* could likewise be seen as discursive commitments, because acts and attitudes belong to different (and mutually exclusive) ontological categories (acts are events or episodes, while attitudes are states or properties). It must then be admitted that commitments (and *a fortiori* doxastic commitments) are not the only kinds of things for which reasons may be asked/given, these also include certain acts or performances.

It could rightly be objected that it is a mistake to look at commitments only as things for which reasons may be asked/given, and that we must also look at the way in which they are involved in the *giving* of reasons, even if this implies restricting our

attention to discursive (and even doxastic) commitments. But what is it to "give" a reason, and what kinds of things are apt to be *given* as reasons, i.e., as making one (either *ceteris paribus* or all things considered) entitled to something?

On the view we are considering, it is fairly clear that to give a reason is *to do* something which makes it manifest that one accepts, and therefore has, a certain doxastic commitment; in Brandom's words, it is to *acknowledge* or endorse a certain doxastic commitment. Yet questions arise (i) as to whether what is thereby being given as a reason is the doxastic commitment itself (one's having this commitment) or merely its content, and (ii) as to whether the reason so given could be a reason for a further doxastic commitment except in virtue of being in the first place a reason for *acknowledging* this commitment. As it happens, Brandom says both that doxastic commitments are what can be given as reasons, and that reasons are what are given as the contents of doxastic commitments. He also seems to hold that at least some reasons may be reasons to *acknowledge* certain doxastic commitments, and hence to perform a certain kind of *act*, which threatens to conflict with his view that inferential relations are relations between deontic statuses, as well as with his claim to have provided an account of intentionality in normative terms (since these *acts* of acknowledgement would remain unaccounted for). These are questions concerning the terms and nature of the relations between reasons and what they are reasons for; and since these are supposed to be closely allied to inferential relations, it will be helpful to turn to Brandom's conception of the latter.

Brandom (1994, p. 168-169) describes inferential relations as relations of inheritance and/or exclusion between (discursive) deontic statuses. In what he calls the intrapersonal (or concomitant) dimension of inferential articulation, he recognizes three kinds of inferential relations. Commitment-preserving inferential relations are such that one cannot be committed to the premises without thereby being committed to the conclusion; entitlement-preserving relations are such that one cannot be entitled to being committed to the premises without being entitled to commitment to the conclusion; and incompatibility relations between two commitments are such that having one of them precludes being *entitled* to having the other, which means that one cannot have the one commitment without thereby *failing* to be entitled to have the other.

What is striking in these characterizations of inferential relations is that they don't involve the concept of a reason at all, and they don't even suggest that there is any normative relation linking either premises to conclusion or commitment/entitlement to the premises to commitment/entitlement to the conclusion. To say, for example, that one cannot be committed to p without being committed to q is just to say that the one commitment somehow "entails" or "necessitates" the other, and doesn't involve any normative relation between these two commitments (let alone the relation of "being a reason for"). Yet it seems that *some* normative relation must be in play somewhere, if *inferences* (as opposed to inferential relations, as characterized above) are to be described as correct/incorrect or appropriate/inappropriate. If the aim is, as I think it is, to account for the inferential articulation of *contents* in normative terms, it doesn't seem to carry us very far to be told, say, that p entails q in virtue of the fact that being committed to p entails being committed to q. This suggests that to find what we are looking for, namely, how commitments get involved in the giving of reasons, we must look at actual *inferences* and not only at (inferential) relations between deontic statuses.

Consider a particular commitment-preserving inference, such as "the sky is red, therefore, it is not blue". This is what Brandom would describe as a materially correct deductive inference. In *making* such an inference, one *acknowledges* being committed to the sky's being red, and takes the fact that it is red (i.e., the *content* of one's commitment) as a reason (in this case a *conclusive* reason) to *acknowledge* being committed to its not being blue. There may be some uncertainty as to exactly how the notion of a conclusive reason relates to that of being entitled "all things considered", but the two notions are certainly very close. There is also some uncertainty as to whether one's conclusive reason to *acknowledge* being committed to the sky's not being blue can or should also be described as one's reason for *being* so committed, since we are here assuming that the agent already is so committed (in virtue of the fact that one cannot be committed to the sky's being red without being committed to its not being blue). But let us suppose there is some derivative sense in which it can. Still, the only normative relations in play here are between a *content* and the acknowledgement of a commitment, and/or between a content and a commitment.

And it is intuitively clear that in making such an inference one is not giving one's being committed to the sky's being red as one's reason to acknowledge being committed to the sky's not being blue. For if this were what one is doing, the inference would not be correct, because being committed to the sky's being red *is no (normative) reason* to acknowledge commitment to the sky's not being blue (and *a fortiori* no reason for being so committed), in either of the two senses that we have considered so far (namely the *pro tanto* sense and the "all things considered" or "conclusive" sense). To see this more clearly it must be reminded that acknowledging a commitment is here understood in such a way that it involves *endorsing* that commitment (and not merely admitting it). Clearly, that one is committed to the sky's being red doesn't make it the case that one has a reason for (is entitled to) endorsing a commitment to the sky's not being blue; for if it did, it would mean that being committed to the sky's being red (or indeed, to anything) is in itself a (good) reason for endorsing it, or in other words that it suffices to make it the case that one is *entitled* to this commitment.

This is not to deny that one may indeed take one's being committed to the sky's being red as one's reason to acknowledge commitment to its not being blue. But this would have to be expressed by saying something like "I am committed to the sky's being red, therefore, the sky is not blue", and would immediately be seen to be incorrect. In other words, to *take* something as one's reason doesn't make it a reason. Moreover, in making such an inference, one is not acknowledging commitment to the sky's being red, but to *being committed* to the sky's being red; in other words, *that one is committed* to the sky's being red appears as the *content* of some further commitment, which does nothing to show that commitments themselves are or can be reasons.

It is no help to observe that being committed to the sky's being red can *of course* be a or the reason *why* one is committed to the sky's not being blue, in virtue of the fact that (as we are assuming) the one commitment entails the other, since the notion of a reason *why* is not normative, and that of entailment, in any case, has not yet been shown to be normative (even though this is part of what the whole project is ultimately aiming at). Hence, it remains quite unclear how commitments, as opposed to their contents, are involved in the giving of reasons.

It might be complained that we are not looking in the right direction. Just as we found the connection between entitlements and reasons by considering the use of the *verb* "to entitle", perhaps we may hope to find the connection between commitments and reasons-giving by considering the use of the *verb* "to commit". But the most natural way to understand the claim that one thing commits one to another, is as saying that the first thing makes it the case that one is committed to the other. And on this construal, it is simply false that being committed to the sky's being red *commits* one to being committed to the sky's not being blue (for this then says that being committed to the sky's being red makes it the case that one is committed to one's being committed to the sky's not being blue, which is intuitively false, or at least very odd).

On the other hand, it seems natural to hold that being committed to the sky's being red *does* make it the case that one is committed to *acknowledging* being committed to the sky's not being blue, and thus, that the first commitment *commits* one (in the intended sense) to *acknowledging* the other. I have already insisted that acknowledgements are not deontic statuses, and this may be a source of trouble since inferential relations are supposed to be relations between deontic statuses. But independently of this, the problem here is that even though this relation certainly appears to be normative in some sense, it remains unclear how its normative character relates to that of reasons. For it was argued above that being committed to the sky's being red is not a reason for acknowledging (endorsing) commitment to its not being blue. Here it may be added that it would still leave us with a puzzle if it were, since it would mean that the normative relation involved in saying that something *commits* one to endorse a certain commitment is not different from the one involved in saying that something *entitles* one to endorse this commitment, which would clearly be unacceptable.

At this point, it must, I think, be recognized that the reason relation (i.e., the relation of "being a reason for") simply is a relation between contents and either commitments or the acknowledging of commitments (or both). So if, as certainly seems to be the case, there is a normative relation between being committed to the sky's being red and acknowledging being committed to its not being blue, it must be of a different sort. And here we may want to make a distinction between being rational and having reasons, i.e., between the

normativity of rationality, which has to do with the relations between one's acknowledgements of commitments, and the normativity of reasons as such. The suggestion, in a nutshell, is that even if being committed to the sky's being red is no reason for acknowledging commitment to its not being blue, it would still be *irrational* for someone committed to the sky's being red not to acknowledge commitment to its not being blue.

This sounds intuitive enough, but notice it would not be acceptable to say that it would be irrational for someone committed to the sky's being red not to be committed to its not being blue. For *that* is not *irrational*, it is plainly impossible; which I take to show that there is still no normative relation between the two commitments, but only a normative relation between one commitment and the *acknowledgement* of another. Now the latter relation can itself plausibly be seen as deriving from the obtaining of a corresponding relation between acknowledging commitment to the sky's being red and acknowledging commitment to its not being blue; or it will at least be granted that it would be just as irrational for someone who acknowledged commitment to the sky's being red not to acknowledge commitment to its not being blue.

The upshot is that if the aim is to account for the fact that the sky's being red entails its not being blue in terms of the *propriety* of inferring the latter from the former, it would seem much more promising to look at normative *rationality* relations involving acknowledgements of commitments than to look at any normative relations between commitments (for there just doesn't seem to be any such normative relations). As we might tentatively put it, it is in virtue of the fact that it would be irrational to acknowledge commitment to the sky's being red while refusing to acknowledge commitment to its not being blue (i) that the sky's being red entails its not being blue, *and* (ii) that the sky's being red is (or would be, if true) a (conclusive) reason for (acknowledging) being committed to its not being blue. We could of course go on to claim that it also is in virtue of this same fact that being committed to the sky's being red entails being committed to its not being blue. But what would be the point of doing so?

Once we have reached this point, it becomes hard to see what real work the concept of commitment is supposed to be doing. It looks as if we could dispense at least with the

(supposedly deontic) status of being committed, and appeal only to the idea that some acknowledgements "rationally commit" one to (or as I will put it below, "rationally require") others. In other words, we seem to be in a position to account for the inferential relations between contents (and perhaps also of the reason relation between contents and discursive commitments) in terms of normative relations between intentional acts (which I have already insisted cannot be deontic statuses).

### 3. *Inference and Normativity*

Let us now have a closer look at the two other kinds of (intrapersonal) inferential relations, and ask how the ambiguity that has been found in the notion of entitlement reflects on Brandom's characterizations of them. This will lead us to unearth a further ambiguity and to revise Brandom's classification of inferential relations. The resulting characterizations of inferential relations will then, hopefully, put us in a position to come back to the normativity issue.

I start with entitlement-preserving relations. The first thing to observe is that in the very same paragraph in which he describes entitlement-preserving relations as being such that one cannot be *entitled* to being committed to the premises without being entitled to commitment to the conclusion, Brandom (1994, p. 169) states that "[t]he premises of these inferences entitle one to commitment to their conclusions [...] but do not compel such commitment. For the possibility of entitlement to commitments incompatible with the conclusion is left open". I think it is fairly clear, intuitively, how this statement is to be understood, yet it is puzzling that it doesn't say anything about being *entitled* to the premises (or about *inheritance* of entitlement).

Consider someone who is committed to a certain match's being dry. *That* this match is dry is a reason for (and thus *prima facie* entitles) such a person to acknowledge commitment to the claim that it will ignite, if struck. But (looking at it from a certain angle) this has nothing to do with this person's being or not being *entitled* to being committed to the match's being dry; it looks as if one could not be committed to the match's being dry without being (*prima facie*) entitled to the claim that it will ignite if struck. In accordance with Brandom's remark that "the possibility of entitlement to commitments

incompatible with the conclusion is left open", the same person could also be committed to the match's being at a very low temperature, and thereby be (*prima facie*) entitled to the claim that it will not ignite if struck.

It might be thought that the condition that one must be *entitled* to (commitment to) the premises in order to be entitled to (commitment to) the conclusion becomes relevant when it is *all things considered* entitlement which is in question. But it is obvious that someone who is committed to the match's being dry, and *all things considered* entitled to this commitment, can still fail to be *all things considered* entitled to the claim that it will ignite if struck. For one can be (committed and) *all things considered* entitled to both the claim that the match is dry and the claim that it is at a very low temperature, and in such a case one will *not* be *all things considered* entitled to the claim that it will ignite if struck (though one will still be *prima facie* entitled to it).

However, when we look at it from another angle, it seems just *incredible* that the mere fact that one is committed (or even acknowledges commitment) to the match's being dry makes it the case that one is even *prima facie* entitled to being committed to its igniting if struck. Such entitlements plainly are too cheap, which seems to justify the requirement that one be *entitled* to being committed to the match's being dry.

What this shows, I think, is that we should have a closer look at the relations between reasons and entitlements. A reason, I said, is something that *entitles* one to acknowledge some (discursive) commitment and which can be given as the content of a (doxastic) commitment. Obviously, a (propositional) content as such could not entitle one to anything unless one is actually (doxastically) committed to it, yet its being a reason for this or that commitment doesn't depend on anyone's being committed to it. Now suppose p is a *pro tanto* reason for (acknowledging) being committed to q. Then there are two senses in which one might say that the fact that p (*prima facie*) entitles one to acknowledge commitment to q. On one way of reading this claim, what it says is that (i) one is committed to p and (ii) one's being so committed *would* make one *prima facie* entitled to q, *if* one were entitled to being committed to p; and on another reading, what it says is that (i) one is both committed to p and entitled to this commitment and (ii) one's being so

committed and entitled makes it the case that one actually<sup>2</sup> is *prima facie* entitled to acknowledge commitment to q. So, there is a distinction to make between what I will call *conditional* and *unconditional* entitlement, which is *not* to be confused with the distinction between *prima facie* and *all things considered* entitlement.

This distinction is not restricted to *pro tanto* reasons. For suppose now that p is a sufficient or conclusive reason for (acknowledging) being committed to q. Then to say that the fact that p (*all things considered*) entitles one to acknowledge commitment to q can be understood either as saying that (i) one is committed to p and (ii) one's being so committed *would* make one *all things considered* entitled to q, *if* one were *all things considered* entitled to being committed to p, or as saying that (i) one is both committed to p and *all things considered* entitled to this commitment and (ii) one's being so committed and entitled makes it the case that one actually is *all things considered* entitled to acknowledge commitment to q.

Moreover, it seems that all (and perhaps only) Brandom's commitment-preserving inferential relations actually belong to the latter category of *all-things-considered*-entitlement-preserving inferential relations (i.e., to what I call below inferential relations of the ATC type). For example, consider again the inference from "the sky is red" to "the sky is not blue". In making this inference, one is giving a conclusive reason to endorse commitment to the sky's not being blue; which means that one's being committed to the sky's being red would make one *all things considered* entitled to being committed to the sky's not being blue, if one were *all things considered* entitled to being committed to the sky's being red.

Thus, with the distinction between conditional and unconditional entitlement relations, both of Brandom's descriptions of entitlement-preserving inferential relations can be seen to be acceptable, provided one is read as involving unconditional entitlement and the other as involving conditional entitlement. Furthermore, depending on whether we read these descriptions as involving *prima facie* or *all things considered* entitlements, we get different kinds of inferential relations, one of which seems to correspond to Brandom's

---

<sup>2</sup> "Actually", but still "subjectively", insofar as it could turn out to be false that p, in which case there is a further sense in which one will fail to have an adequate reason to endorse q.

"commitment-preserving" inferential relations, thus suggesting that there is no need for a special category of commitment-preserving inferential relations.

I have pointed out that some entitlement-preserving inferential relations are such that *prima facie* entitlement to the premises makes one *prima facie* entitled to the conclusion (let's call them inferential relations of type PF), and that some others are such that *all things considered* entitlement to the premises makes one *all things considered* entitled to the conclusion (type ATC). But it seems also intuitively clear that some entitlement-preserving inferential relations will be such that *all things considered* entitlement to the premises makes one *prima facie* (but not *all things considered*) entitled to the conclusion (type ATC-PF), and that *no* inferential relation will be such that *prima facie* entitlement to the premises makes one *all things considered* entitled to the conclusion (type PF-ATC). It seems just as clear that all inferential relations of type ATC-PF also belong to type PF, and that all inferential relations of type ATC also belong to type ATC-PF and to type PF. The interesting question I don't know how to answer is whether all inferential relations of type PF also belong to type ATC-PF.

Let us now briefly consider Brandom's third kind of inferential relations, namely incompatibility relations, which Brandom explains (1994, p. 169) by saying that p and q are incompatible propositions when commitment to p precludes entitlement to q. It is fairly clear, in light of the previous discussion, that this can only be understood as involving conditional entitlement, i.e., as saying that one's being committed to p would make it the case that one is entitled to (acknowledge) commitment to q if one were entitled to commitment to p. For otherwise (as I have already pointed out), the mere fact that one has two mutually incompatible commitments would prevent one from being entitled to *any* of them, which is certainly not what is intended. Hence, incompatibility relations turn out to be entitlement-exclusion relations (instead of relations between commitments and entitlements).

And now the question arises what kinds of entitlement-exclusion relations there are. Clearly, there must be entitlement-exclusion relations of type ATC, i.e., such that one's being committed and *all things considered* entitled to p precludes one's being *all things considered* entitled to q (which is arguably the same as making it the case that one is *all*

*things considered* entitled (to acknowledge) not to be committed to q). But as far as I can see, and somewhat surprisingly, there doesn't seem to be entitlement-exclusion relations of any other kind. There are no entitlement-exclusion relations of type PF-ATC, i.e., such that one's being committed and *prima facie* entitled to p precludes one's being *all things considered* entitled to q, no entitlement-exclusion relations of type ATC-PF, i.e., such that one's being committed and *all things considered* entitled to p precludes one's being *prima facie* entitled to q, and no entitlement-exclusion relations of type PF, i.e., such that one's being committed and *prima facie* entitled to p precludes one's being *prima facie* entitled to q. However, I take this to be an anomaly, and to indicate that we have overlooked something.

To appreciate what it is that is missing, we must return to the idea of a *pro tanto* reason, and properly register the fact that there are both positive *and* negative reasons, reasons for *and* reasons against. For example, the fact that this match is at a very low temperature is a reason against (acknowledging) being committed to its igniting if struck. Suppose one is committed and *prima facie* entitled to this match's being at a very low temperature; it would seem that one is thereby *prima facie* entitled to refrain from (acknowledging) being committed to this match's igniting if struck, or as we might put it, that one is thereby *prima facie* dis-entitled to (acknowledge) being committed to the claim that this match will ignite if struck. It should be obvious, however, that this doesn't mean that one could not also be *prima facie* entitled to (acknowledge) being committed to the match's igniting if struck (e.g., in virtue of the fact that one is also committed and *prima facie* entitled to this match's being dry), for there is no incoherence in having both a reason for and a reason against one and the same thing. Just as one can be *prima facie* entitled to both p and not-p, one can also be both *prima facie* entitled and *prima facie* dis-entitled to p. Hence, there is after all some sort of "negative" inferential relation between the claim that this match is at a very low temperature and the claim that it will ignite if struck: an entitlement-repelling relation of type PF.

It should now be easy to see that Brandom's entitlement-exclusion relations really are entitlement-repelling relations of type ATC. One's being committed and *all things considered* entitled to the sky's being red makes it the case that one is *all things considered*

dis-entitled to (acknowledge) being committed to the sky's being blue. Of course, it also *precludes* one's being *all things considered* entitled to (acknowledge) being committed to the sky's being blue, but this is only because one cannot be both *all things considered* entitled and *all things considered* dis-entitled to one and the same thing. Clearly, there must also be entitlement-repelling relations of type ATC-PF, if there are entitlement-repelling relations of type ATC, since one cannot be *all things considered* dis-entitled to p without being also *prima facie* dis-entitled to p. But there can be no entitlement-repelling relations of type PF-ATC, i.e., such that one's being committed and *prima facie* entitled to p makes it the case that one is *all things considered* dis-entitled to q.

The upshot is that all (intrapersonal) inferential relations are either entitlement-preserving or entitlement-repelling. This looks like an improvement, but on the other hand, it is still unclear how any of this could help to see inferential relations as being grounded in normative relations between commitments and/or entitlements. As far as I can see, the conclusion still stands, that the relevant normative relations basically involve the acknowledgements of commitments rather than commitments or entitlements themselves.

Let us consider again the two basic types of entitlement-preserving inferential relations, namely type ATC and type PF. Suppose one is committed and *all things considered* entitled to the sky's being red. As a result, one is thereby *all things considered* entitled both to be committed to the sky's not being blue and to acknowledge being so committed; but this is not in virtue of there being any normative relation between being committed and *all things considered* entitled to the sky's being red and being *all things considered* entitled to be committed to the sky's not being blue and to acknowledge being so committed. So far, there are only consequential relations between entitlements. Yet it seems such a relation could easily be introduced, in the following way. Let us say that being committed to p "rationally requires" acknowledging commitment to q iff one could not be committed and *all things considered* entitled to p without being *all things considered* entitled to q. Turning now to PF type entitlement-preserving relations, we could say, in much the same way, that being committed to p "rationally recommends"<sup>3</sup> (or "rationally supports") acknowledging commitment to q iff one could not be committed and *prima facie*

---

<sup>3</sup> I borrow these terms from Broome (1999).

entitled to p without being *prima facie* entitled to q. Proceeding in similar fashion, one could thus introduce normative relations corresponding to each type of entitlement-preserving or entitlement-repelling inferential relations.

The point is this: once these normative relations have been made available, there is (as far as I can see) nothing to prevent one from reversing the perspective and take them as the primitives in terms of which the consequential relations between entitlements (or between entitlements and dis-entitlements), and ultimately all inferential relations among contents, are to be explained.

One more point. The normative "rationality" relations I have just alluded to are relations between a commitment and the acknowledgement of a commitment (or, taking entitlement-repelling relations into account, between a commitment and a "dis-acknowledgement"). However, one might not be completely satisfied with such relations. For suppose again that one is committed and *all things considered* entitled to the sky's being red, but this time suppose further that one doesn't (and/or wouldn't) acknowledge being so committed (perhaps because one's commitment to this claim is a remote consequence of one's other commitments). It isn't clear in such a case that one is rationally required to acknowledge commitment to the sky's not being blue, nor that one would necessarily be rational if one were to acknowledge being so committed (since one could acknowledge this commitment on other, incorrect, grounds). Such worries might lead us to prefer saying that *acknowledging* commitment to p "rationally requires" acknowledging commitment to q if and only if one could not be committed and *all things considered* entitled to p without being *all things considered* entitled to q; and similarly for the other relevant normative relations. These, I think, are two possible and legitimate ways to go, though they rest on different intuitions about the force and nature of the norms of rationality.

All of this remains somewhat incomplete and sketchy, but I think what I have said can easily be seen to point towards two main conclusions.

The first is, as announced at the beginning, that we do not have a firm grasp of the normative concepts of commitment and entitlement in terms of which Brandom frames his account, and hence do not understand exactly how the latter is supposed to work. I have

urged that the only way to see the real normative significance of these concepts would be by making it explicit exactly how they relate to the more familiar, but almost equally elusive, concepts of "reason" and "rationality".

The second, and more tentative, conclusion is that if there is to be any hope of providing an explanation of the inferential articulation of contents in normative terms, one must start with appropriate normative *relations* (instead of statuses), and give pride of place to acknowledgements (i.e., to intentional *acts*, as opposed to statuses) as terms of these relations, which could therefore not themselves be explained in terms of deontic statuses.<sup>4</sup>

**Daniel Laurier**

*Université de Montréal*

daniel.laurier@umontreal.ca

---

<sup>4</sup> A version of this paper has been read at UQAM on March 31<sup>st</sup> 2007. I am grateful to Robert Brandom for his sympathetic comments, and to Josée Brunet, for the numerous discussions that we have had on these topics.

## References

- Brandom, Robert B. (1994) *Making it Explicit*. Cambridge: Harvard U. Press.
- \_\_\_\_\_. (2001) 'Modality, Normativity, and Intentionality', *Philosophy and Phenomenological Research* 63, 587-609.
- \_\_\_\_\_. (2008) *Between Saying and Doing: Towards an Analytic Pragmatism*, Oxford, Oxford U. Press.
- Broome, John. (1999) 'Normative Requirements'. In Jonathan Dancy (Ed.) *Normativity* (pp. 78-99). Oxford: Blackwell.
- \_\_\_\_\_. (2004) 'Reasons'. In R. Jay Wallace, P. Pettit, S. Scheffler and M. Smith (Eds.) *Reason and Value*. Oxford: Oxford University Press, pp. 28-55.
- Millar, Alan. (2004) *Understanding People: Normativity and Rationalizing Explanation*, Oxford, Oxford U. Press.
- Rosen, Gideon. (2001) 'Brandom on Modality, Normativity and Intentionality', *Philosophy and Phenomenological Research* 63, 611-623.
- Sosa, Ernest & Villanueva, Enrique (Eds.) (2005) *Philosophical Issues 15: Normativity*. Oxford: Blackwell.
- White, Heath. (2003) 'Brandom on Practical Reason', *The Philosophical Quarterly* 53, 566-572.

## ON THE INTERPRETATION OF HUME'S EPISTEMOLOGY

João Paulo Monteiro

### Abstract

At the end of his life, Hume neglected his first work, and declared that he wished his readers to take into account only the later versions of his theories of the understanding, the passions and morals. This poses a special problem of interpretation: is there a difference between a "young Hume" and a "mature Hume", as in the case of Hegel, and several other thinkers? Is there in Hume's work anything comparable with the shift from the pre-critical to the critical Kant? I believe that Hume's case does not fall in any of these categories, but that it still poses problems analogous at least to the first, that is, Hegel's. This is the hypothesis this essay aims to investigate in the particular case of Hume's epistemology. I defend the view that a correct interpretation of Hume's epistemology only becomes possible after a careful reading of his more mature works. I illustrate this by discussing Hume's distinction between association by causation, on the one hand, and causal inference on the other, as well as his concept of experience.

Hume's work is a unique case in the history of philosophy. He left us a first book which he never republished, choosing instead to rewrite it in smaller ones with the same basic content, but with large differences under several aspects. At the end of his life he declared that he never acknowledged that first work, and that he wished his readers to take into account only the definitive versions of his theories of the understanding, of the passions and of morals. This poses a special problem of interpretation: is there a difference between a "young Hume" and a "mature Hume", as in the case of Hegel, and several other thinkers, as the non-philosopher Karl Marx? Is there in Hume's work anything comparable with the shift from the pre-critical to the critical Kant? I believe that Hume's case does not fall in any of these categories, but that it still poses problems analogous at least to the first, that is, the Hegel/Marx situation.

Of course, the main difference is, to restrict ourselves to the subject of the understanding, that in Book I of the *Treatise of Human Nature* the young Hume presents a theory of knowledge that is fundamentally the same as the one we find in the first *Enquiry*. Disappointed with the reception of the first work, he wrote twelve essays in which we find some of the problems discussed in it, intitled, in 1748, *Philosophical Essays concerning*

*Human Understanding*, a title that would be replaced, in the 1751 edition, by another title suggesting a unified work: *An Enquiry concerning Human Understanding*. I shall not discuss here the possible intentions of our author when he made these choices, but shall only repeat what many have remarked before me: if the *Enquiry concerning Human Understanding* was first called *Philosophical Essays*, it must be because the author did not intend it to be a unified version of the clearly unified exposition of Book 1 of the *Treatise*.

But those *Essays* cannot be reduced to something like a set of "selections" from Book 1 of the *Treatise*. On the other hand, only sections 2 to 7 present a consistent new version of Hume's theory of the understanding, to which we must add sections 11 and 12. Section 1 does not correspond to the Introduction to the *Treatise*, but consists in a completely new kind of preface, if we choose to call it so. Section 8 corresponds to themes of the book on the Passions, and sections 10 and 11 are entirely new. We know there was a first discussion of the problem of miracles, which the author chose to eliminate. But section 11, on the problems of theology and teleology is new, and more akin to the later *Dialogues concerning Natural Religion* than to any subject in the *Treatise*.

Ten sections of the *Enquiry* correspond to Book 1 of the *Treatise*, and any reader is able to find in the first, among other things, a shorter version of Hume's theory of the understanding. Why then did its author, in his famous Advertisement to the last edition of his *Essays and Treatises on Several Subjects*, express his wish that everybody refrained from regarding the *Treatise* "as containing his philosophical sentiments and principles"?<sup>1</sup> He presents some fragmentary explanations of his decision: "that juvenile work, which the author never acknowledged", had gone "to the press too early", and presented several "negligences" in its reasoning and in its expression.

One negligence that was probably due to haste in "going to the press" concerns the form of presentation of the third principle of association of ideas. After commenting on resemblance and contiguity, Hume adds: "As to the connection, that is made by the relation of *cause and effect*, we shall have occasion afterwards to examine it to the bottom, and

---

<sup>1</sup> *An Enquiry concerning Human Understanding* (EHU), Tom Beauchamp, ed., Oxford University Press, Oxford, 1999, p. 83.

therefore shall not at present insist upon it."<sup>2</sup> Now, when Hume turns again to that kind of relation, he asserts that, contrary to resemblance and contiguity, it "is requisite to persuade us of any real existence",<sup>3</sup> which only makes sense if we take that relation to be more than a mere relation of association. It only makes sense in case it consists in the stronger relation in which are founded "all reasonings concerning matter of fact",<sup>4</sup> that is, concerning "real existence".<sup>5</sup>

Now, this is the beginning of Hume's argument about *causal inference*, an argument where there is no place for the subject of association, and this is part of Hume's final and definitive version of his epistemological theory. In the light of this theory, the third principle of association is not significantly stronger than the two others, and its discussion in the *Treatise* may receive a better interpretation if it is suspected, and perhaps recognized, that in the beginning of Book I Hume was not yet able to make a clear distinction between association by causation, on the one hand, and causal inference on the other, a distinction which is crystal clear in the *Enquiry*.<sup>6</sup> A coherent, if not true (for that, perhaps, would be too much to be hoped for...) interpretation of Hume's epistemology becomes possible only if we admit, not only that its definitive and correct version is that of the *Enquiry*, but also that the *Treatise* is guilty of mistakes that are perhaps "more in the manner than the matter", as Hume wrote in one of his letters, but that, as Dorothy Coleman once said in a Hume Conference in S. Paulo, are "more in the manner, but also in the matter."

The negligences in the *Treatise* affect the problem of the interpretation of Hume in a richer way than if that problem consisted simply in the coexistence of two different versions of the same theory, the first of them being disavowed by the author. We may imagine several possible attitudes. The first is the most common, and consists in ignoring the problem, studying and teaching Hume's epistemology as if there was a perfect compatibility between them, approaching each particular subject using the method of

---

<sup>2</sup> *A Treatise of Human Nature* (THN) 1.1.4, David Fate Norton and Mary J. Norton, eds., Oxford University Press, Oxford, 2000, p. 13.

<sup>3</sup> THN 1.3.9, p. 76.

<sup>4</sup> EHU 4.1, p. 109.

<sup>5</sup> *Ibid.*, p. 108.

<sup>6</sup> See my *Novos Estudos Humeanos*, Chapter 1, Discurso Editorial, S. Paulo, 2003, pp. 15 ff., where further examples of Hume's "negligences" concerning association and causation are discussed.

"indifferent quotation" that is, indiscriminately picking passages in both works in order to discuss that subject. This has been, I must confess, my own method for many years. It should be clear by now that, at least concerning the relation between association and causal inference, this method deserves to be revised. We may even wonder whether Hume, during the composition of the *Treatise*, had any clear notion of the difference between causal inference and association by causation: I was unable to find in that work a single example of association by cause and effect. On the other hand, we find in the more mature, definitive version of his epistemology a clear example of this kind of association: (...) "if we think of a wound, we can scarcely forbear reflecting on the pain which follows it." And in a note Hume gives this an indubitable name: "Cause and effect", (EHU 3.3), that is, unequivocally, association of ideas by cause and effect, or by causation. And it is evident that this relation between wounds and pains is not an example of causal reasoning... We shall see below that another example of association by cause and effect was given in the *Abstract*.

On the other hand, the *Treatise* may be considered the greatest explosion of philosophical genius to occur in the first half of the XVIIIth century, and the comparison between Book 1 and the first *Enquiry* shows that, if the latter is clearly a corrected version, containing a more perfect philosophy of knowlewdge, as Hume himself more than suggests, the first included several themes and developments whose suppression we must regret. We miss the clarity and scope of the Introduction, as well as of the distinction between memory and imagination in 1.3 and the enumeration of the philosophical relations in 1.5, even though we are apt to feel differently about the absence of the Lockean themes of modes and substances. We may also regret the loss of the development of the subject of space and time in 2.1 to 6, although the rather obscure section 7 about the problem of existence was perhaps mercifully suppressed.

In Part 3 we have the bewildering addition, as a *fourth* principle of association of ideas, of something like "association by repetition" (in 3.14), followed by the strange contention that this "true principle of association among ideas" is "the very same with that between the ideas of cause and effect, and (...) an essential part in all our reasonings from that relation" (3.15). In my chapter on these problems, quoted in note 6 above, I hope to

have shown that no reasoning directly depends on this or any other principle of association, but only on the influence of custom or habit on our imagination, as the *Enquiry* strongly maintains. But we must read 3.15 on "other habits" to fully understand that the Humean concept of custom is much stricter than the common one. Also, the subject of probability has been too strictly contracted in EHU 6, and we ought to miss sections 11 to 13 in the *Treatise* (1.3). Also, Section 15 on general rules, whose importance has been rightly emphasized in Fred Wilson's *Hume's Defence of Causal Inference*,<sup>7</sup> must be carefully read in order to understand some important aspects of Hume's general epistemology.

Finally, Part 4, on several forms of scepticism and on two delicate subjects, the immateriality of the soul and personal identity, presents us a different set of problems. Beginning with the last: Hume himself declares in a letter his dissatisfaction with his own treatment of the problem of personal identity, but the persistence of the subject through time, with dozens of Hume scholars writing on it, may lead us to suspect that he was too harsh in his judgment on himself, and that he suppressed that section for reasons still to be discovered. Suppression of the section on the soul may find a satisfactory explanation in Hume's writing of the essay with the resemblant title "Of the Immortality of the Soul", although the two texts differ in several important respects. But the discussion of scepticism, although reduced in the new version to some pages of the *Enquiry* (12), is a simpler version that may help us to understand, as I believe it should, that Hume was not a sceptic of any other kind but his own particular brand: a special form of "academical scepticism" totally new in his time, which may be taken as an important step towards Peirce's fallibilism and, perhaps, of most "ex-analytical" philosophers of the 20<sup>th</sup> century.

All this leaves us in a rather complicated situation. If Hume's advertisement was meant to lead us to neglect the *Treatise*, we clearly would have to say "no", with all due respect. Not only would that be a loss to our philosophical culture, but we wouldn't be able to discuss important points like (one among many) the question of "other habits". But if he meant to lead us to give a strong priority to the *Enquiry*, ignoring all passages in the *Treatise* that are incompatible with it, and abstaining to argue from any part of that "juvenile work" against Hume himself, or against the interpretations that may be suggested,

---

<sup>7</sup> Fred Wilson, *Hume's Defence of Causal Inference*, Toronto: University of Toronto Press, 1997.

in the *first* place, only by the *Enquiry*, in that case we should entirely assent to the suggestion of our philosopher. Priority of the mature version, but no global rejection of the younger Hume: this might be a provisional guide for our interpretation of his epistemology.

But one cannot help rejecting some passages in the *Treatise*, in the light of Hume's mature epistemology, as we can find it the *Enquiry*. The very concept of experience is a case in point. We read in that juvenile work:

The nature of experience is this. We remember to have had frequent instances of the existence of one species of objects; and also remember, that the individuals of another species of objects have always attended them, and have existed in a regular order of contiguity and succession with regard to them. Thus we remember to have seen that species of object we call *flame*, and to have felt that species of sensation we call *heat*. We likewise call to mind their constant conjunction in all past instances. Without any farther ceremony, we call the one *cause*, and the other *effect*, and infer the existence of the one from that of the other.<sup>8</sup>

Hume then describes this as the discovery of constant conjunction as a "new relation" between cause and effect. Now, the only "negligence" we may detect in this otherwise impeccable version of his theory of inference is that it implies that the nature of experience includes *repetition*, under the form of constant conjunction, when even in the same work, a dozen pages after this definition of experience, we see that even for the younger Hume what is essential to have experience is simply to have *conjunction* of kinds of phenomena, even without repetition.

"(...) not only in philosophy, but even in common life, we may attain the knowledge of a particular cause merely by one experiment, provided it be made with judgment, and after a careful removal of all foreign and superfluous circumstances."<sup>9</sup> This becomes possible "when we have lived any time",<sup>10</sup> which means that repetition is part of "the nature of experience" only ... when we have had only very little experience. After that, if one wants to define the nature of experience one should say that experience occurs when we observe *conjunctions* of phenomena, either repeatedly or in "single experiments". In the absence of any conjunction, Hume generally speaks of simple *survey*, not of experience,

---

<sup>8</sup> THN 1.3.6, p. 61.

<sup>9</sup> THN 1.3.8, p. 73. I discuss this subject in my *Novos Estudos Humeanos*, ed. cit., Chapter 3, pp. 65 ff.

<sup>10</sup> EHU 9.5, note, p. 167.

although he never explicitly established this difference in vocabulary. But Hume's epistemology certainly isn't, as we may see when our own reading of it is not ... negligent, anything like "a slave of repetition" ... This is a case when a possible misinterpretation may be avoided even if we resort only to the *Treatise*, but even here the *Enquiry*, as we have seen, helps to clarify such an important subject as that of the nature of experience in Hume's philosophy.

The problem of Hume's particular kind of scepticism is quite intractable in the *Treatise* (no pun intended), not so much because of any negligences properly so called, but simply due to a certain imprecision in the statement of the philosopher's position towards Pyrrhonism and other forms of scepticism. The inspiration for Popkin's assignment of a kind of Pyrrhonian scepticism to Hume derives from certain vague phrases in that juvenile work, like for instance the following:

The intense view of the manifold contradictions and imperfections of human reason has so wrought upon me, and heated my brain, that I am ready to reject all belief and reasoning, and can look upon no opinion as more probable or likely than another.<sup>11</sup>

Now, it is almost understandable that Popkin, commenting on this same passage, jumps to such conclusions as the following:

A close examination of Hume's views will show that he agreed with the Pyrrhonian theory of the inability to find any rational and certain basis for our judgments (...); we have no ultimate criterion for determining which of our conflicting judgments in certain fundamental areas of human knowledge are true, or to be preferred.<sup>12</sup>

The spirit of the *Treatise* does ambiguously seem to authorize such interpretations. But even Popkin's moderate thesis (he also insists in Hume's critique of Pyrrhonism, as we see in the very title of his paper), is corrected by Hume's definition of scepticism in the *Enquiry*:

---

<sup>11</sup> THN 1.4.7, p. 175.

<sup>12</sup> Richard Popkin, "David Hume: his Pyrrhonism and his critique of Pyrrhonism", in V. C. Chappell, ed., *Hume*, Macmillan, London, 1970, pp. 56-7.

(...) scepticism, when more moderate, may be understood in a very reasonable sense, and is a necessary preparative to the study of philosophy, by preserving a proper impartiality in our judgments (...) To begin with clear and self-evident principles, to advance by timorous and sure steps, to review frequently our conclusions, and examine accurately all their consequences; though by these means we shall make both a slow and short progress in our systems; are the only methods, by which we can ever hope to reach truth, and attain a proper stability and certainty in our determinations.<sup>13</sup>

I think it is needless to insist in Hume's reference, in his mature philosophy, to truth and certainty; even in a sceptical key, Popkin's interpretation cannot make sense of this philosophy. The mature Hume was what he clearly says that he was: a mitigated sceptic (or, again, a kind of fallibilist), not any other kind of sceptic.

In all three cases examined here, the latter passage seems to contradict the first. And I believe it is sounder to accept that it really does, expressing a deep change in Hume's epistemology, than to accept what Flew called "the Infallibility Assumption", which consists in "insisting that where two passages in an author appear to be inconsistent, one of these passages has to be so interpreted that the apparent inconsistency is resolved". Flew ridicules this assumption, adding that it should never be confused with

the entirely sound and proper rule that we should always employ all the resources of scholarship in the attempt to show, what may of course turn out not to be true, that any apparent absurdities or apparent inconsistencies in our author are when properly understood neither absurdities nor inconsistencies.<sup>14</sup>

I agree with Flew's position. In the problems of interpretation examined in this paper, I think that we should not adopt any dogma of infallibility, at the same time that the second rule, although it is quite reasonable in itself, is equally improper to be applied here. Instead, I believe we should reflect on the problems I mentioned first, about the peculiar character of Hume's work: in this work taken as a whole there is no sharp difference, like in Kant or Hegel, between the two versions of his philosophy, to whose epistemological aspect I restrict myself here, but there still are important differences, and these should be taken seriously and examined with the utmost care.

---

<sup>13</sup> EHU 12.1.4, p. 200.

<sup>14</sup> Antony Flew, "On the Interpretation of Hume", V. C. Chappell, ed., *Hume*, ed. cit., p 280. This paper, whose title obviously inspired mine, is about some problems in Hume's moral philosophy.

In the interpretation proposed above I only tried to show some significant differences between the young and the mature Hume. But there is still room for further questions, mainly about *why* there are such differences, between works after all relatively similar in content like those examined here. To these second-order questions only one standard interpretation may suffice, mainly that Hume simply noticed and corrected some of his mistakes or negligences. But maybe more than one interpretation is in order, perhaps a different one in each case, to account for what Noxon considered to have been our philosopher's "philosophical development".<sup>15</sup>

In the first place, the *Treatise* reveals a desire to explain causal inferences in terms of association. This produced a muddle that remains as one of the main negligences in that work, as it seems to me to be clear enough. Much less clear is perhaps the exact nature of the motivation that led Hume to insist on that untenable explanation in the framework of his thought, first giving us the impression that he did not clearly distinguish between association by causation and causal inference, as we have seen above, and secondly, as I have tried to show in my *New Studies*, introducing in the *Treatise*, some dozens pages after his enumeration of only three principles of association of ideas, an ambiguously presented fourth principle of association of ideas, which we may call "association by repetition".<sup>16</sup>

Why would Hume resort, in part 3 of Book 1, to a Lockean concept of association, incompatible with the three ones, in part Aristotelian, that he had introduced in Part 1? From this moment on, all possible interpretations we may propose must be by far more speculative and uncertain than the precedent ones. But, with this in mind, we may perhaps dare to notice that in the Introduction to the same work Hume defends that in the science of man, like Newton in natural science, "we must endeavour to render our principles as universal as possible, and explaining all effects from the simplest and fewest causes."<sup>17</sup>

Could Hume have been unable to resist to the temptation of explaining human knowledge by one principle instead of several ones, and elect association for that central role? We may notice that, when he first presented his concept of association, he famously said that those "principles of union or cohesion among our simples ideas" are comparable to

---

<sup>15</sup> James Noxon, *Hume's Philosophical Development*, Clarendon Press, Oxford, 1973.

<sup>16</sup> See the first chapter of my book, pp. 24 ff.

<sup>17</sup> THN Introduction, p. 5.

"a kind of Attraction, which in the mental world will be found to have as extraordinary effects as in the natural, and to show itself in as many and as various forms."<sup>18</sup> Newton had explained a great variety of phenomena by gravitational attraction; could Hume feel tempted to try to explain the human mind by only one principle, namely, association? We know that he couldn't achieve this, and he also knew this at least in 1748, when he published an *Enquiry* where association and habit, at least, are principles of comparable importance, and both are indispensable to explain the phenomena of knowledge and many others, thus proposing an explanation where not one, but two principles or more, concur in the production of mental phenomena.

Another case in point is that of the *Abstract*, an anonymous pamphlet where in 1740 Hume tried to present Book 1 of his *Treatise* to the general public. The most developed subject is that of causation, but in the last paragraph Hume adds that, among the "new discoveries in philosophy" presented in that work, "if anything can entitle the author to so glorious a name as that of an *inventor*, it is the use he makes of the principle of association of ideas, which enters into most of his philosophy."<sup>19</sup> There is a vast difference between these high ambitions concerning the scope of association in Hume's epistemology and the role to which association is reduced in the *Enquiry*, a role already mentioned in the *Treatise*: connecting ideas in the imagination, giving it a certain regularity, and binding simple ideas in complex ones.<sup>20</sup> And also the secondary role of serving as an *illustration* of the principle of transition of vivacity from impressions that is responsible for the production of the lively ideas that we call "beliefs".<sup>21</sup> That is, by the time of the publication of the *Treatise* our philosopher apparently thought that his principles of association could have a role comparable to Newtonian universal attraction. But from 1748 on he never allows these pious hope to be revived, choosing instead, in all the eight editions the *Enquiry* had until his death, to give the central role in the production of causal reasoning to custom or habit, a principle that has nothing to do with association of ideas.

---

<sup>18</sup> THN 1.1.4, p. 14.

<sup>19</sup> David Hume, *An Abstract of a Book lately published, entitled A Treatise of Human Nature, etc.*, which was included at the end of the Norton edition of THN. The quotation is from p. 416, in which we also find the association between father and son as a (first) example of association by cause and effect.

<sup>20</sup> EHU 3.1, p. 101.

<sup>21</sup> EHU 5.2, pp. 126-9.

Hume's careless definition of experience only by repetition in the *Treatise* (1.3.6) quoted above also conflicts with one of the most important passages in the *Enquiry*. That definition suggests, to say the least, that our typical road to the acquisition of empirical knowledge consists in having repeated experiences of conjunctions. But in the *Enquiry* (5.1.5) we find a new face of Hume's epistemology, more theoretical than empiricist, when we see that his theory about the discovery of causal relations by repeated experiences is not an empirical description of everybody's inferences, but exclusively concerns only what we may call a "primeval subject", the theoretical invention of an imaginary knowing subject. This subject is a person "endowed with the strongest faculties of reason and reflection", but one who never had any kind of experience, for she has been "brought on a sudden into this world", and must have some repeated experiences before she can reach her first causal conclusions.<sup>22</sup> It is for this kind of theoretical being, whom we obviously can never meet in real life, that "the nature of experience" can be understood in terms of repeated experience alone. One should perhaps speak of a kind of "mitigated empiricism" as the mark of Hume's epistemology. This mitigated empiricism has a double face. First, the real knowing subject is not supposed to need repeated experience to make causal discoveries, these being possible also starting from a single observation of one conjunction followed by a relatively complex and partly deductive inference. And second, while experience surely is the condition of all knowledge of the world, Hume's own theory never consists in conclusions derived from observation of our verifiable cognitive behaviour, but is supported by the theoretical invention of the primeval subject, a being who is not empirically accessible. Of course, in his first work Hume had not discovered this, whence the air of "simple empiricism" that pervades that work. A correct interpretation of Hume's philosophy, here as elsewhere, only becomes possible after a careful reading of his more mature works.

Hume's scepticism, as we have seen, also cannot be rightly interpreted unless we, not only make a careful reading of his definitive epistemology, but also go through the pains of an even more careful reading of the juvenile work where our philosopher may sometimes seem to have fallen in some kind of radical or pyrrhonian scepticism. Maybe he hesitates, or maybe he is guilty of some negligences, as he himself admits. But we, as

---

<sup>22</sup> EHU 5.1.3, p. 120.

readers who want to do justice to the greatest philosopher of the English language, who was also perhaps the greatest philosopher of the eighteenth century, should never allow ourselves any negligence in the study of Hume, like for instance opting for the easy "method" of presenting all his texts without a clear distinction between the less careful work which he wrote in his youth and his more solid and definitive philosophy.

Problems concerning the interpretation of Hume's philosophy, either the epistemology discussed here, or the moral, metaphysical or political aspects of his work, shall always be open to discussion and criticism. No particular version can aspire to achieve general agreement. I only hope that every problem and every passage in Hume's work receives a careful and impartial examination from Hume scholars in general. For a fruitful discussion, perhaps each one could indicate which possible findings in Hume's writings would lead her to change at least one of her cherished interpretations. For my part, if I could be shown, in Hume's mature works, any clear defence of associationism about causation, or of common empiricism, or of anything equivalent to pyrrhonism, I would gladly change my views about Hume's epistemology.

**João Paulo Monteiro**

*Universidade de São Paulo*

jpmonteiro@netcabo.pt

## References

- Flew, Antony. (1970) 'On the Interpretation of Hume', In V. C. Chappell, ed., *Hume*, Macmillan, London.
- Hume, David. (1999) *An Enquiry concerning Human Understanding*, Tom Beauchamp, ed., Oxford University Press, Oxford.
- \_\_\_\_\_. (2000) *A Treatise of Human Nature*, David Fate Norton and Mary J. Norton, eds., Oxford University Press, Oxford.
- \_\_\_\_\_. (2000) 'An Abstract of a Book lately published, entitled A Treatise of Human Nature, etc.', In *A Treatise of Human Nature*, D.F. Norton and M.J. Norton, eds., Oxford University Press, Oxford.
- Monteiro, J.P. (2003) *Novos Estudos Humeanos*, Discurso Editorial, S. Paulo.
- Noxon, James. (1973) *Hume's Philosophical Development*, Clarendon Press, Oxford.
- Popkin, Richard. (1970) 'David Hume: his Pyrrhonism and his critique of Pyrrhonism', In V. C. Chappell, ed., *Hume*, Macmillan, London.
- Wilson, Fred. (1997) *Hume's Defence of Causal Inference*, Toronto: University of Toronto Press.

**TOPICS IN PHILOSOPHY OF LANGUAGE, MIND AND SCIENCE:  
PROCEEDINGS OF THE SECOND EUROPEAN GRADUATE SCHOOL**

**Editorial**

**Second European Graduate School: Philosophy of Language, Mind and Science**

The following three graduate student articles were selected from among nineteen first-rate presentations which were presented during the “Second European Graduate School: Philosophy of Language, Mind and Science”, organized at Ruhr-University Bochum and the University of Lausanne in March 2009. We received fifty high quality graduate submissions from students of European as well as overseas universities. The submissions were without exception subjected to a double-blind review process, and we would like to take this opportunity to thank all our colleagues for their valuable assistance in this time-consuming reviewing process.

The Graduate School was the result of a collaboration financed by a DAAD program between the philosophy departments at Bochum (Germany), Lausanne (Switzerland) and Tilburg (The Netherlands), who had decided to coordinate their graduate education in philosophy of language, mind and science. To this end, two week-long meetings were organized in order to allow selected graduate students to present and discuss their ongoing research projects. Each week focussed on one main topic, which was discussed in extended tutorials by the two keynote speakers. A one-day international workshop with several invited speakers rounded off each week’s program.

The first of these two weeks took place in Bochum and dealt with the topic “Self, Person, and Action”. This part of the workshop was combined with the Carnap Lectures, an event taking place annually since 2008 at Ruhr-University Bochum. This year’s lectures were given by John Perry (Stanford University), who focused on several aspects of the self. François Recanati (Institut Jean Nicod, Paris), our main speaker for the Graduate School, discussed central aspects of context-dependency. We would like to thank John and François, but also the graduate students and other international speakers participating in the

workshop for stimulating discussions, which took place in an informal and cheerful atmosphere.

The second week in Lausanne was centered on the topic “The Philosophy of Perception”, with Tim Crane (University of Cambridge) and Michael Tye (University of Texas at Austin) as keynote speakers. Both gave several lectures that dealt with hotly debated issues in the philosophy of mind and perception. We would like to take this opportunity to express our gratitude to Michael and Tim for their stimulating talks as well as for the ensuing discussions that took place in a very friendly atmosphere to the great benefit of all participants. In addition, we should also like to thank the invited speakers who presented talks during the Graduate School’s closing workshop. They all contributed significantly to the success of the Lauanne week, making this a memorable event.

This two-week long event marked the second instalment of a program of three European Graduate Schools. The third meeting will take place in Lausanne and Tilburg in October 2010, and we are confident that it will be just as successful as its two predecessors.

The selected papers offer inventive proposals on three quite diverse issues. 1. How to conceive of semantic reference in natural languages? 2. What are the prospects for an intentionalist theory of self-deception? 3. What account of perceptual consciousness do synaesthetic experiences call for?

As to the first paper, Jessica Pepp (University of California, Los Angeles) compares two conceptions of semantic reference. The *conventional* conception is contrasted with what is coined the *historical* conception of semantic reference. It is argued that the two conceptions are both ways of conceiving of *semantic* reference, and that the historical conception is more viable as a basis for the semantics of natural language than the conventional conception. The paper finishes by drawing a distinction between a *theory* of semantic reference and the historical *conception* of semantic reference, describing the latter as setting the stage for the former.

The second paper dwells upon the claim that most or all self-deceptions depend on intentional self-deception. Kevin Lynch (Warwick University) argues that intentional models of self-deception can partly be traced to a particular invalid method for analyzing reflexive expressions of the form ‘Ving oneself’ (where *V* stands for a verb). In addition, it

is argued that the best prospects for an intentionalist theory of self-deception lie with a strategy involving the control of attention.

Finally, the third paper by Michael Sollberger (University of Lausanne) addresses the issue of what an indirect realist theory of perception should look like. More precisely, the goal of his article is to prompt a new view of perceptual consciousness that is ruthlessly structural. To this end, he combines the structural approach to representation with an original discussion of empirical cases of synaesthesia. He challenges our intuitions by arguing that there are good reasons to conceive of some synaesthetic experiences not as illusory or hallucinatory, but as truly veridical perceptions. In addition, he highlights in his contribution how synaesthetic experiences are well-suited to corroborating a structural account of the perceptual mind.

Last but not least, special thanks are due to the editors of ABSTRACTA, who enabled us to make the outstanding graduate papers assembled in this volume accessible for a wide range of readers.

**Albert Newen**

*Ruhr-Universität Bochum*

albert.newen@rub.de

**Raphael van Riel**

*Ruhr-Universität Bochum*

raphael.vanriel@rub.de

**Michael Sollberger**

*Université de Lausanne*

michael.sollberger.2@unil.ch

## SEMANTIC REFERENCE NOT BY CONVENTION? <sup>1</sup>

Jessica Pepp

### Abstract

The aim of this paper is to approach a basic question in semantics: what is semantic reference? Or, what is reference, insofar as the notion has a role in the semantics of natural language? I highlight two ways of conceiving of semantic reference, which offer different starting points for answering the question. One of these conceptions – what I call the *conventional* conception of semantic reference – is the standard conception. I propose an alternative to this conception: what I call the *historical* conception of semantic reference. The first section of the paper explains the two conceptions, highlighting their common ground and how they differ. The second section offers a preliminary argument that the two conceptions are really both ways of conceiving of *semantic* reference, and that the historical conception is more viable as a basis for the semantics of natural language than the conventional conception. Finally, in the third section, I comment on the status of the historical conception as a basic view about semantic reference that sets the stage for (but does not constitute) the development of a *theory* of semantic reference.

### 1. Two conceptions of semantic reference

To present the two conceptions of semantic reference perspicuously, it will be useful, first, to lay out common ground between the two conceptions. This common ground involves many of the notions I will be relying upon, so discussion of it will serve to introduce these notions, as well. There are three points of agreement among the two conceptions that I would like to highlight.

First, proponents of both conceptions can agree that *utterances* of expressions have *historical explanations*. I mean this only in the very broad sense in which there is some kind of story to tell about what gives rise to a given event, such as an utterance.

---

<sup>1</sup> This paper was originally published, in a somewhat different form, under the title “Two Conceptions of Semantic Reference,” in *Meaning, Content and Argument. Proceedings of the ILLI International Workshop on Semantics, Pragmatics and Rhetoric*, Jesus M. Larrazabal and Larraitz Zubeldia, eds., University of the Basque Country Press, 2009. I would like to thank the editors for granting me permission to publish a modified version of the paper here. This paper has been improved by the comments and suggestions of a number of people. I benefited in particular from the comments of attendees of the Second European Graduate School in Bochum, Germany, the Spring 2009 conference of the University of Iowa Graduate Philosophical Society, the ILLI International Workshop on Semantics, Pragmatics, and Rhetoric, the UCLA Language Workshop, and a writing workshop at UCLA. I would also like to thank Joseph Almog, Antonio Capuano, Eliot Michaelson, Terry Parsons, Andrew Reisner, and an anonymous referee for this journal, and especially Brendan Gillon for extensive written comments and discussion.

Second, it should be agreed that *expressions* of language have *conventions of use*. For instance, it is a convention of English that “I” is used to refer to oneself. No one can reasonably dispute this.

The third thing that should be agreed upon is that, at least in many cases, the conventions of language will suggest what might be called “conventional referents”. Here is how. Any convention must apply to something. It is outside the scope of this paper to discuss the theory of convention, but perhaps it is safe to say that a convention applies to a type of situation. For instance, if one is served a bowl of soup, it is conventional that one eat the soup with a spoon. If one is introduced to a new person, it is conventional to shake the new person’s hand.<sup>2</sup> And so on. Conventions for using referring expressions must also fit this pattern: it must be possible to describe the type of situation in which it is conventional to use a given expression. For instance, one might describe the type of situation in which it is conventional to use “I” in the following way: if one intends to make oneself the subject of discourse, use “I”. Or, if one stands in a certain kind of historical relation to oneself, use “I”. Or, most simply, use “I” if you are referring to yourself.<sup>3</sup>

Given that conventions for referring apply to situations in which someone is referring to something, it is possible to abstract from them, at least in some cases (empty names might be an exception), the *conventional referent* of an expression (or of an occurrence of that expression). This is the individual that the speaker *would* be referring to *if* she were using the expression in accord with convention.

This much, I am assuming, is common ground. There are utterances, which have historical explanations, and there are linguistic expressions, which are associated with conventions. Because these conventions apply to situations in which speakers are referring to things, we can speak of the “conventional referents” of expressions as those things that the speaker would be referring to, if she were using the expression in accord with convention.

With these points as background, I will now characterize the two conceptions of semantic reference, noting the points on which they differ. On the conventional conception

---

<sup>2</sup> Of course, such conventions vary by culture and geography.

<sup>3</sup> Note that this introduces the idea of non-conventional notion of referring, such that a convention for referring is a convention of using an expression to refer, in this non-conventional way, to a certain thing. This is central to the discussion in section 2, below.

of semantic reference, referring expressions in a language are associated with certain conventions, which determine their semantic referents, perhaps relative to a context of use.<sup>4</sup> The way I am using it here, “determine” means “make to be the case”. So on the conventional conception, the convention associated with an expression, perhaps relative to a context, makes the expression’s referent be its referent. This is in contrast with the use of “determine” on which it means “figure out”, or “reveal”. The conventional conception does not hold that a referring expression, as used in a given context, has a referent already, independent of the convention, which the convention in some way reveals. Rather, the convention makes the referent of the expression be what it is.

Now it is true that on the conventional conception, referring expressions have come to be associated with their conventions via the ongoing processes of language formation and change. But at any given time in the history of a language, there is a convention as to how any referring expression of that language refers. When a speaker uses a referring expression, the semantic referent of her use of the expression is determined by the convention, regardless of the history of that particular use of the expression. It is not that the conventional conception of reference ignores history. Conventions arise in linguistic communities over time, and thus have histories. But on the conventional conception, the history of the convention is pre-semantic. For instance, there is a history behind the convention governing my use of “I”. However, when I use “I”, the current convention that “I” is used to refer to oneself simply applies as it stands, and *determines* that this occurrence of “I” semantically refers to me, Jessica Pepp.

A third point is that the conventional conception holds that what makes an expression be a referring expression is the fact that there is a convention of using it to refer. What makes “I” a referring expression is that there is a convention associated with it whereby it is used to refer to whoever uses it.

Each of these points is in contrast to the historical conception of semantic reference. The basis for the historical conception is the relation between an utterance of a referring expression and that which gave rise to the utterance. Another way to put it is that on the

---

<sup>4</sup> For instance, in David Kaplan’s semantics for indexicals, linguistic convention supplies a “character” for the expression “I”, which determines that the referent of “I” relative to a given context of use is the agent of that context; Kaplan (1989).

historical conception, the semantic referent of an expression as uttered on a given occasion is part of the historical explanation of how that expression came to be uttered.<sup>5</sup> To use a phrase due to Joseph Almog, an expression as used in a given utterance has a referent,  $x$ , because  $x$  is the “source of a chain” leading to the use of the expression in that utterance.<sup>6</sup> Thus, semantic reference itself is a historical relation between an expression as uttered and the referent of that uttered expression. The conventions associated with a referring expression do not determine – in the sense of “make to be” – its referent (relative to a context). They may help to determine – in the sense now of “reveal” or “figure out” – what the referent of the expression as uttered is. But they do not make it have a certain referent.

So one important difference between the two conceptions is that on the historical conception, referring expressions have semantic reference only relative to utterances. Indeed, what makes an expression be a semantically referring expression is not the existence of a convention of using it to refer, but the fact that a particular utterance of it has been generated in a certain way.

Before moving on, let me offer an analogy to make the historical conception more vivid. The historical conception views the relation of semantic reference as analogous with the ownership relation between an email address and its owner. Suppose that a company, Corputech, Inc., has a policy that each employee is to have an email address of the form firstname.lastname@corputech.com. When you receive an email from someone with the sender address: theodore.thomas@corputech.com, knowledge of the Corputech convention may lead you to guess that the owner of that email address is the Corputech employee named Theodore Thomas. But you will also be aware that this might not be the name of the employee who sent the order. There might have been an error in setting up the address, and the sender's name may actually be “Thomas Theodore,” or “Theodora Thomas”, or something completely different. There may be no Corputech employee at all by the name “Theodore Thomas”. And even if there was no error, it is clear that the Corputech convention

---

<sup>5</sup> I find this idea primarily in the work of Keith Donnellan on referential uses of definite descriptions (1966) and empty names (1974). Of course, neither of these phenomena is my subject in the present paper, but I think my notion of a historical conception of semantic reference is closely related to Donnellan’s “historical explanation theory” of reference.

<sup>6</sup> Almog (2004: 404-405).

does not determine (in the sense of “make to be the case”) that the address theodore.thomas@corputech.com is owned by the employee by that name. Rather, it was the work done by the information technology specialist in setting up the address that made it the case that this person has this address. Your knowledge of the Corputech email assignment convention is something that can help you figure out who the sender is, but the convention itself does not *make* the referent be what it is. The convention is a guide, not a determiner.

Just as the owner of an email address is part of the historical explanation of how that email address was set up, so, on the historical conception, the semantic referent of an expression is part of the historical explanation of how that expression came to be used.

It should be noted that the historical conception of semantic reference does not and should not deny that language users exploit conventions of language to aid in communication, or even that such exploitation of conventions is required for large portions of our communication. The historical conception only denies that conventions of language are determinative of semantic reference; that what it is for an expression to semantically refer to something is for it to refer to it by convention.

To sum up: the two conceptions of semantic reference agree that there are expressions, conventions associated with expressions, utterances of expressions, and historical explanations of utterances. They differ in that the conventional conception takes expressions to have semantic referents determined by conventions associated with those expressions, regardless of the historical explanations of particular utterances; whereas the historical conception takes expressions to have semantic referents only relative to utterances of those expressions, where the semantic referents are part of the historical explanations of those utterances.

## **2. Non-conventionality of semantic reference**

The historical conception of semantic reference may seem like a category error. Utterances are speech acts, one might say; things we do with a language that already, independently, has its semantics. In making utterances we may capitalize on the semantic interpretation of a language or we may flout it, but that is irrelevant to what the semantic interpretation of

the language is. Expressions of a language mean what they mean – and have the semantic referents they have – as provided for by the conventions of the language, independent of what any speaker may have in mind, and independent of what led to a speaker's use of an expression on a given occasion.<sup>7</sup> Thus, when understood correctly, the historical and conventional conceptions are compatible: it is just that they are not both conceptions of *semantic* reference. The historical conception is of some other, non-semantic relation.

This line of thinking can be challenged by reflecting on how referring conventions might arise in natural language. One view of how referring conventions might arise is suggested by Saul Kripke's influential account of name reference. On Kripke's view, a name refers to something in virtue of a convention having been passed along a "chain of communication", from user to user.<sup>8</sup> When I use a name, it refers to something in virtue of my having taken on a convention of using it to refer to that thing - the convention passed to me by the person from who I acquired the name. This convention may have been instituted originally by what Kripke calls a "baptism": an event in which someone fixes the referent of an expression by stipulating that it will refer to a particular thing. According to Kripke, this is whatever fits the description used by the baptizer. For instance, in Kripke's famous example, Leverrier introduces "Neptune" as a name for whatever satisfies the description, "the cause of the perturbations in the orbit of Uranus".<sup>9</sup>

But whether a given thing satisfies this description depends on what, for instance, "Uranus" refers to. Presumably, a view like Kripke's will say that the reference of "Uranus" in Leverrier's description is similarly determined by the convention taken on by Leverrier when he acquired the name "Uranus". And that convention will have similarly been instituted by some description-involving baptism (or reference-fixing), the satisfier of which

---

<sup>7</sup> See, for example, Kripke's (1977: 263) critique of Donnellan (1966), in which Kripke operates on the assumption that the notion of reference relevant to semantics is a matter of the conventions of the language.

<sup>8</sup> It might seem that Kripke conceives of reference historically in the way I have described. However, I think it is clear from his discussion that the historical relation in Kripke's account of reference is between a speaker's acquisition of a name and the initial introduction of a convention of using the name to refer to a given thing (the "baptism"). The reference relation itself is a conventional one, given by the convention introduced. The historical relation in the account is between the expression and its associated convention, not between the expression and its referent.

<sup>9</sup> Kripke (1980: 79, footnote 33).

will depend on what some other expressions refer to, which will depend on some other description-involving introduction, and so on.

To stop this regress, reference-fixing descriptions must ultimately be grounded in expressions whose reference had not been fixed by a description, but in some other way. Or, the Kripkean “baptismal” story might be disregarded at the outset, and one could argue that referring conventions are established without such reference-fixing events. On either approach, there is appeal to what might be called a “brute convention.” No reference-fixing descriptions or baptisms are involved: a convention of using some name *N* to refer to some individual *x* just arises. This convention can then be passed from speaker to speaker as they acquire the expression from one another. Having acquired *N* and entered into this conventional practice of using it to refer to *x*, when one now uses *N*, the convention determines that the use of *N* refers to *x*.

But consider what it is for a convention of using *N* to refer to *x* to “just arise”. In a stipulative reference-fixing event like a Kripkean baptism, the baptizer *mentions* the expression and stipulates that it will refer to whatever fits a certain description. With this convention established, subsequent *uses* of the expression would refer to that thing. However, if a convention arises simply because speakers *use* *N* to refer to *x*, then these pre-conventional uses of *N* are just that: *uses* of *N*, not *mentions* of *N*. As uses, they must be interpreted. Suppose someone is struck by the strangeness of another person she sees, and declares, “Garsaloosius walks among us,” making up the name “Garsaloosius” because it seems to suit the strange appearance of that other person. This is an introduction of the expression “Garsaloosius”, but it is also a use of the expression. “Garsaloosius” is not introduced as an uninterpreted sign, which will become interpreted if a convention arises of using it to refer to the strange looking person. Indeed, the idea that a convention for referring could arise just by virtue of people using the expression to refer to something depends on the pre-conventional uses being of an interpreted expression - an expression that refers. If the initial uses of the name were uninterpreted, they would not provide the basis for the development of a convention for using “Garsaloosius” to refer to something.<sup>10</sup>

---

<sup>10</sup> Note that in the present discussion, I am mentioning - not using - the expression “Garsaloosius”.

It is commonplace to note that the project of semantics for natural languages deals with *interpreted* languages. The aim is to understand the interpretation of a language, not to specify it. Thus, in characterizing semantic reference, one should not ignore the fact that referential expressions, if they are used, are interpreted, regardless of whether any convention is associated with them. This suggests that convention cannot be the foundation of semantic reference. The historical conception of semantic reference that I have outlined seems a better account of the nature of semantic reference.

There are responses to be made on behalf of the conventional conception of semantic reference. I will briefly consider two. I do not have definitive replies to them, but I will explain why I do not think either will work. First, one might respond by insisting that pre-conventional uses of an expression (for instance, the initial use of “Garsaloosius” in my example) are not uses of interpreted expressions. They are speech acts which allow the *speaker* to be interpreted (i.e., we can say what *she* refers to, but not what her expression refers to). The expression used only comes to be interpreted once a convention for using it to refer in this way arises.

I do not think this response works because “Garsaloosius” is a linguistic expression (it is not, for instance, an inarticulate grunt), and it is *used* in the example I gave. In the example, the expression is introduced by being used. This was crucial for avoiding the regress problem associated with introductions by reference-fixing stipulation. A linguistic expression cannot be used (as opposed to mentioned) without being interpreted. And the business of natural language semantics is to understand the interpretation of natural language. Thus, it seems to me that “Garsaloosius” in its initial use has semantic reference if anything does.

Another response on behalf of the conventionalist would be to accept that there is historical semantic reference in initial cases, but that once a convention arises it replaces the historical relationship as the semantic relationship. Thus, in the initial use of “Garsaloosius”, the semantic referent might be whatever plays the appropriate role in the historical explanation of the use. However, once the convention has arisen, the semantic referent of the expression is what the convention determines.

This introduces an odd disunity in the nature of semantic reference, however. In pre-conventional uses, an expression has as its semantic referent something figuring in the historical explanation of its use. Later on, once a convention is associated with the expression, its semantic referent is determined by that convention, even though there are still historical explanations of uses of the expression. It is not that this is an impossible view to hold, but it does strike me as *ad hoc*, designed to make semantic reference as much as possible a conventional relation, with exceptions for the cases where this does not seem plausible.

### 3. Status of the historical conception

I think these considerations tell against a conventional conception of semantic reference. This does not mean that there are no worries for the historical conception. Obviously, much is left to be articulated about this conception. I have not given an account of precisely what kind of historical relation reference is. If the task of developing the historical conception of semantic reference is thought of as providing a guide to finding referents for what Donnellan called an “omniscient observer of history” (hereafter, “OOH”), then the guide I have offered might seem hopelessly vague. All it says is that the OOH should pick out something that is in some way historically related to the utterance. I have not said in what way. So how has any account of reference been provided? It is not open to me to reply that *I* may not know what constitutes reference, but the OOH does. That would make the account of reference uninteresting, because we would be saying that reference is just whatever relation a being who knows everything would say it is.

But I do not think the prospects for the historical conception of semantic reference are so bleak. The situation can be viewed as analogous with the case of seeing. In times before vision science, people probably believed that seeing was a historical relation seers had to the things they saw. That is, they probably believed that the things they saw in some way gave rise to their seeing them. They probably did *not* believe that seeing involved having one's experience just happen to match up in some way or other to the thing seen.

Of course, prior to vision science, the nature of the historical relation between seers and things seen was unknown. Similarly, the nature of the historical relation between an expression and its referent needs more investigation. But just as the starting point of the

investigation of the relation of seeing was the basic view of it as a historical relation, so a starting point for the investigation of the relation of reference is the basic view of *it* as a historical relation. The historical conception of semantic reference is not so obviously correct as is the historical conception of seeing, but it is a view at the same - basic - level. The difficulties of working out a historical *theory* of reference do not invalidate the historical *conception* of reference.

**Jessica Pepp**

*University of California, Los Angeles*

pepp@humnet.ucla.edu

## References

- Almog, J. (2004) 'The proper form of semantics', IN A. Bezuidenhout and M. Reimer (eds.), *Descriptions and Beyond*, Oxford: Oxford University Press, pp. 390-419.
- Donnellan, K. (1966) 'Reference and Definite Descriptions', *The Philosophical Review* 75, 281-304.
- \_\_\_\_\_. (1974) 'Speaking of Nothing', *The Philosophical Review* 83, 3-31.
- Kaplan, D. (1989) 'Demonstratives', IN J. Almog, J. Perry, and H. Wettstein (eds.) *Themes from Kaplan*, Oxford: Oxford University Press, pp. 481-563.
- Kripke, S. (1977) 'Speaker's Reference and Semantic Reference', *Midwest Studies in Philosophy* 2, 255-276.
- \_\_\_\_\_. (1980) *Naming and Necessity*, Cambridge, Massachusetts: Harvard University Press.

## PROSPECTS FOR AN INTENTIONALIST THEORY OF SELF-DECEPTION

Kevin Lynch

### Abstract

A distinction can be made between those who think that self-deception is frequently intentional and those who don't. I argue that the idea that self-deception has to be intentional can be partly traced to a particular invalid method for analyzing reflexive expressions of the form 'Ving oneself' (where *V* stands for a verb). However, I take the question of whether intentional self-deception is possible to be intrinsically interesting, and investigate the prospects for such an alleged possibility. Various potential strategies of intentional self-deception are examined in relation to Alfred Mele's suggestion that doing something intentionally implies doing it knowingly. It is suggested that the best prospects for an intentionalist theory of self-deception lie with a strategy involving the control of attention.

### 1. Two approaches to the analysis of self-deception

The self-deception debate is riven by a theoretical divide between so-called 'traditionalists' and 'deflationists' (though many philosophers take up mixed positions between them). These theoretical differences can be traced in large part to two different approaches to how the concept of self-deception should be properly analyzed, usefully distinguished by Alfred Mele. One, he calls the *lexical* approach, the other, the empirical or *example-based* approach. On the lexical approach; '[w]e might start by asking what "deception" (or "deceive") means, and then ask what "self-deception" must mean if it is to be a species of deception'.<sup>1</sup> Alternatively, on the example-based approach; '[o]ne starts by gathering and constructing cases that would generally be described as self-deception, and then attempts to develop an analysis of self-deception on the basis of a consideration of this material. The meaning of "self-deception" is determined by the cases, which are therefore the most fundamental data'.<sup>2</sup> Note that these two approaches are somewhat idealized, and it can be difficult to find a philosopher who explicitly and exclusively adopts one.

---

<sup>1</sup> Mele (1987: 13).

<sup>2</sup> Mele (1987: 13-14). After making this two-fold distinction in a 1987 paper, he later introduced another category; the 'theory-guided approach'. As he defines it, on this approach 'the search for a definition is guided by commonsense theory about the etiology and nature of self-deception (1997: 92). However, Mele

Both approaches yield quite different answers to the question of what self-deception must be. Mele himself practices the example-based approach. He thinks that the meaning of a term of folk-psychology is a function of how ordinary folk use it.<sup>3</sup> Accordingly, he starts by considering typical cases in which we would pre-theoretically refer to someone as having deceived him/herself: the terminally ill patient in denial, the husband who won't believe that her wife is having an affair when it should be obvious, the mother who won't believe that her son is taking drugs, etc. What goes on in such cases, Mele argues quite persuasively, is that judgment becomes distorted and biased by desire and emotion in ways that the subject *didn't intend*.

On the lexical approach, the characterization is different. Basically, this approach starts by establishing a definition of deception from the interpersonal case, and uses it to deduce the meaning of 'self-deception'. Therefore, the meaning of 'self-deception' can be established, on this view, *independently* of looking at or taking into account how ordinary people actually use the expression 'self-deception', and thus independently of any study of the 'garden-variety cases' that Mele speaks of (that the lexical approach may alienate the philosopher from the actual use of this word will be made clearer shortly).

I think that we can understand this approach as being guided, whether explicitly or implicitly, by the following formula. We can call it the 'lexical formula':

What it means for someone to deceive himself is for him to do the same thing to himself that he does to another when he deceives another.

To many an ear this formula may sound intuitively compelling. And it seems to give us the right result in many instances that spring to mind. For example, if Jones shoots Smith, Jones points a loaded gun at Smith and pulls the trigger. And what else, in that case, could it be for Jones to shoot himself, if not to do that same thing, but to himself, namely; point a loaded gun at himself and pull the trigger?

---

doesn't say anything about what these 'commonsense theories' are, and it is unclear what philosophical accounts he has in mind as exemplifying this approach.

<sup>3</sup> Mele (1998: 39).

On the lexical approach, before we can employ this formula to see what it is to deceive oneself, we must first establish what it is to deceive another. The following definition is usually taken to capture what this is:

When A deceives B, A intentionally/deliberately causes B to believe something that A knows/suspects is false.

It is true that some clever counterexamples have been advanced against this definition, e.g. Barnes.<sup>4</sup> However, almost all the philosophers, including Mele,<sup>5</sup> agree that this definition captures the conceptually central or stereotypical cases of interpersonal-deception. So the traditionalist does have scope to argue that the counterexamples are conceptually peripheral or limiting cases of deception, and for that reason she can adhere to this definition using the qualifier 'paradigmatically'. Then, feeding this definition into our formula, we can deduce that paradigmatically:

When A deceives himself, A intentionally/deliberately causes himself to believe something he knows/suspects is false.

Now if we assume the validity of the lexical formula, and of the definition of deception derived from the interpersonal paradigms, then it follows logically that self-deception must be as this definition says. On this picture of self-deception, A, after encountering evidence that makes him realize that some unwelcome proposition  $p$  is true, deliberately causes himself to believe the contrary, welcome proposition not- $p$  (perhaps to avoid the anxiety of knowing that  $p$ ). Some traditionalists argue that we also get the result that A ends up in a condition where he believes that  $p$  and believes that not- $p$  simultaneously. However, it's not obvious why the lexical derivation as it stands would necessarily imply this. Traditionalists here typically advert to the fact that as deceiver, A must believe that  $p$ , and as deceived he must believe that not- $p$ , but this only begs the question of why the person must satisfy both

---

<sup>4</sup> Barnes (1997: 8-11).

<sup>5</sup> Mele (1997: 92).

these roles simultaneously, rather than consecutively.<sup>6</sup> It may be that the traditionalist needs some additional assumptions added to this ‘lexical’ analysis to support this controversial aspect of his/her account.

Nevertheless, on the whole it seems that if we grant the lexical approach, then there is a pretty strong case to be made for the thesis that self-deception must be as this definition states, which by-and-large amounts to the classic traditionalist account. However, it would not follow from this that this phenomenon actually exists. On the lexical approach, the philosopher asks *hypothetically*; what conditions would *have to* be met for there to be a case of self-deception. For that reason, there is room for the question of whether self-deception ever obtains at all or whether it is even possible, and as Mele points out, many who take the lexical approach end up being skeptics about self-deception.<sup>7</sup> Note that this question is ruled out from the outset by the Mele-type approach, since Mele’s methodology takes the legitimacy of people’s customary use of ‘self-deception’ for granted and just asks what goes on in the cases so referred to.

## 2. Problems with the lexical approach

As I’ve said, if we grant the lexical approach we get a strong case for the traditionalist account. However, there appear to be difficulties with this methodology. The problem lies with the lexical formula. We should expect that we could turn this formula into a general formula for deriving the meaning of any reflexive construction grammatically analogous to ‘deceiving yourself’. Accordingly, we can state the lexical formula in general terms as follows:

What it means for one to *V* oneself is for one to do the same thing to oneself that one does to another when one *Vs* another.

---

<sup>6</sup> Though one might point out that in the interpersonal cases A knows that *p* when B acquires the belief that not-*p*, though typical, this is not a necessary element of interpersonal deception. For instance, A could send a deceptive letter to B and die while it’s in transit (see Siegler 1963: 35).

<sup>7</sup> Mele (1997: 92).

...where  $V$  stands for some verb. But as T.S Champlin has shown,<sup>8</sup> such an approach can be parodied for a number of such reflexive constructions.

Consider how the lexicalist reasoning would work for ‘teaching yourself’:

- Teaching yourself is doing the same thing to yourself that you do to another when you teach another.
- If A teaches B about  $x$ , A knows about  $x$  and imparts this knowledge to B.
- Therefore, if A teaches himself about  $x$ , A knows about  $x$  and imparts this knowledge to himself.

Note that our use of the lexical formula has landed us with a ‘paradox of self-teaching’ analogous to the notorious ‘paradox of self-deception’, for it seems as though the one who teaches himself must, as teacher, know about  $x$  and at the same time, as student, be ignorant of  $x$ .

As regards analyzing the notion of being self-taught, the different consequences that would ensue from adopting the lexical approach, as opposed to the example-based approach, are clear and striking. Adopting the former, we get the outlandish result that teaching yourself is imparting your own knowledge to yourself. But this answer clearly doesn’t tally with the actual use of the expression. This use is simply not constrained by the strictures of any such formula. The cases we refer to with that phrase are not cases in which people impart knowledge to themselves (if that is intelligible). They are cases in which people *lack* the relevant knowledge, but acquire it by solitary study, trial and error, and so on, without the benefit of a teacher. Adopting the example-based approach would give an answer like that to the question of what it is to be self-taught. Adopting the lexical approach, we would probably end up being skeptics about the possibility of teaching oneself, or we might construct elaborate metaphysical theories to explain how it is possible to impart knowledge to yourself (mental ‘divisions’ and ‘sub-systems’, etc.), all of which would be a surreal reflection of the self-deception debate.

---

<sup>8</sup> Champlin (1977: 284-285. 1988: 24-25).

It also seems clear that the lexical approach gives us the *wrong* answer for this case. For nobody wants to say that in our ordinary use of that expression we are *misusing* it, because we are referring to cases that are not cases of imparting knowledge to oneself. Therefore, the application of the lexical formula to the case of self-teaching constitutes a *reductio ad absurdum* of the lexical approach, conceived of as a general approach to the analysis of constructions of the form ‘self-*V*’.

Nevertheless, despite the fact that it seems to be the offspring of an invalid analytic method, the traditionalist definition of self-deception has created an interesting philosophical research-program into the bounds of the possible, since some philosophers claim that if self-deception were as this definition states, then deceiving yourself would turn out to be impossible to do (at least under normal circumstances), while others claim that not only is this phenomenon possible, but it’s a common occurrence. Therefore—despite the shortcomings of the lexical approach—I want to *suppose* the lexicalist definition of self-deception, *just for the sake of argument*, to see whether it gets us a possible phenomenon. We can do this so long as we bear in mind that the investigation may have little relevance to our understanding ordinary self-deception, just as the lexical analysis of self-teaching has little relevance for our understanding of ordinary self-teaching.<sup>9</sup>

### 3. The obstacle to intentionally deceiving oneself

Skepticism may arise over this notion of self-deception because of the requirement for it to be intentional or deliberate (terms which philosophers often use interchangeably here). The worry is well-put by Mele:

It is often held that doing something intentionally entails doing it knowingly. If that is so, and if *deceiving* is by definition an intentional activity, then one who deceives oneself does so *knowingly*. But knowingly deceiving oneself into believing that *p* would require knowing that what one is getting oneself to believe is false. How can that knowledge fail to undermine the very project of deceiving oneself?<sup>10</sup>

---

<sup>9</sup> Traditionalists do, however, have other resources for arguing that ordinary self-deception is often intentional, including arguments to the best explanation.

<sup>10</sup> Mele (1997: 92).

Mele's presupposition is that if one Veds intentionally, one necessarily must have *known* or *been aware* that one was doing *that*, namely; Ving.<sup>11</sup> This seems to suggest that 'unconsciously yet intentionally Ving' is contradictory. Let's call this the *knowledge condition* (though we could equally well call it the 'awareness condition'). Philosophers have frequently thought that there's a conceptual connection between the notions of intentional action and knowledge/awareness of what you're doing.<sup>12</sup> Some lexicographers apparently assume so too. For instance, the *American Heritage Dictionary*, 4<sup>th</sup> edition, defines 'deliberate' as 'done with or marked by full consciousness of the nature and effects; intentional'. I assume the relevant considerations that might support such claims would be as follows. Take any occasion when you do something without knowing or being aware that you are doing it. Say, for instance, that you are making soup, and you unwittingly add some sugar (you think that it's salt). Or imagine that you travel to an exotic country with very different customs and you meet one of the locals. You do something that would not attract any notice in your country but which is taken to be highly insulting in this culture, causing great offence to the local, though you weren't aware that this action is considered offensive. Now it seems clear that in these cases we added the sugar or offended the local neither intentionally nor deliberately, and in justifying this, we naturally advert to the fact that we weren't aware that we were doing that. And it's difficult to see how the accusation that we did these things intentionally/deliberately could stand given this lack of awareness. A philosopher could then argue that by parity with such cases, if someone deceived herself intentionally, she must have known that she was doing that.

Though the connection between intention and knowledge/awareness seems to be intimate, there is no opportunity here to investigate whether it allows for exceptions to the knowledge condition. I am just going to grant this condition from here on in. Let's just say that the onus seems to be on the philosopher who assumes that the intentional self-deceiver

---

<sup>11</sup> Actually, whether Mele is really committed to this is unclear, since he states elsewhere that 'hidden intentions' are possible (1997: 100). Nevertheless, Mele does think that there is something paradoxical about the idea of intentionally deceiving yourself, and it's hard to see what else could be generating this paradox if not this presupposition.

<sup>12</sup> See Anscombe (1957: 11 & 87), Miller (1980: 334), Gustafson (1975: 89), Bratman (1984: 387), Hampshire (1970: 145), Donnellan (1963: 406), Moran (2001: 125) and Hamlyn (1971: 46).

is not aware of what she's doing to explain how this is possible.<sup>13</sup> So how exactly does the knowledge condition constitute an obstacle to intentional self-deception?

Let's consider a number of strategies—strategies that could be used to deceive another—and see if they could be used to deceive oneself. Strategies of self-deception can for our purposes be defined as methods for bringing about *deviant* doxastic changes in ourselves, where 'deviant' is supposed to exclude the legitimate ways in which we may bring about belief-changes in ourselves (e.g. acquiring crucial new evidence, noting a mistake in one's reasoning, etc.). Typically, philosophers are interested in self-reliant strategies that may be available for use in ordinary circumstances. So strategies such as hiring a hypnotist or a brain-scientist to accomplish the goal, though possibly effective, are usually not considered interesting (perhaps because they are not contenders for explaining what's happening in ordinary cases of self-deception). Examples of such strategies often mentioned in the literature include the following:

- 1) Lying to yourself
- 2) Manipulating and distorting the evidence
- 3) Rationalizing
- 4) Selective gathering of evidence

Let me consider (1) and (2) first. As I understand it, Mele's argument would take the following form: 'If you lie to yourself intentionally, then (assuming the knowledge condition), you know that *that* is what you are doing, namely, uttering a *lie* (i.e. an untruth). For you to know that would presumably render you immune to it.' Likewise, 'If you distorted the evidence intentionally, then you are aware that you have done *that*, namely, distorted the evidence. Therefore, you couldn't be taken in by it.'

Strategy (3) is trickier. Whether this type of argument works will entirely depend, I believe, on whether 'rationalize' is pejorative. On Kent Bach's definition, rationalization is 'any case of a person's explaining away what he would normally regard as adequate

---

<sup>13</sup> Too often, unconsciously intentional action is attributed to the self-deceiver more in an *ad hoc* manner to save the theory of intentional self-deception, than on the basis of any independent considerations (e.g. Steffen 1986, p.47).

evidence for a certain proposition'.<sup>14</sup> Let's assume here that 'explaining away' is pejorative, so that it means something like 'giving bogus arguments' to undermine evidence. Assuming this, if Jones rationalizes intentionally, and if rationalizing means adducing bogus arguments, then Jones intentionally adduces bogus arguments. And if doing that intentionally implies doing it knowing that *that* is what one is doing, then he *understands* those arguments to be *bogus*. Therefore, he couldn't be taken in by them. Though someone may intentionally rationalize to fool another, the idea of intentionally rationalizing to fool *yourself*—given our assumptions—is incoherent. One would expect rationalizers not to understand themselves to be rationalizing, and so not to be deliberately rationalizing.

Similar arguments apply to (4). To accuse someone of *selectively* gathering evidence is to accuse someone of having done something unjust. To know that one gathered evidence selectively is to know that one did not show justice to both sides of the argument, neglecting one of the sides, and so is to know the dubious value of one's efforts. So granting the knowledge condition, self-deceivers cannot intentionally/deliberately gather evidence selectively.

So we see the form of Mele's argument. A pejorative term is used to denote some epistemically untoward activity. The fact that the strategy is executed intentionally, granting the knowledge condition, implies that the agent knows that his activity is untoward. This knowledge would then preclude the agent being fooled by the product of this activity.<sup>15</sup>

#### **4. The attentional strategy**

However, there are theories of self-deception which mention certain kind of actions that may be intentionally executed, and which may be thought to have potential as an effective means for deceiving oneself, *even granting* the knowledge condition. Robert Lockie

---

<sup>14</sup> Bach (1981: 358).

<sup>15</sup> On some accounts of the strategy of intentional self-deception, the self-deceiver might intentionally do something designed to deceive his/her future self. Here it would be the knowledge of what one *did*, rather than of what one *is doing*, that would be the obstacle to success.

usefully categorizes this type of account under the label ‘attentional accounts’.<sup>16</sup> We can formulate the idea as follows:

A person who believes or at least suspects that  $p$ , but who wishes to believe that not- $p$ , by turning his attention away from the unwelcome considerations supportive of  $p$ , and by attending to welcome considerations supportive of the contrary not- $p$ , may end up losing the (conscious) belief that  $p$  and acquiring the belief that not- $p$ .

So self-deception, on this account, involves what psychologists call *thought-suppression*, i.e. the act of ridding thoughts from the mind—as well as selective focusing of attention. The idea may be illustrated with the following. Harry is a goal-keeper who has lately been letting in some easy shots. He’s beginning to think that he’s not a very good goal-keeper and this distresses him. The attentional theory states that by avoiding the thoughts of his poor performances and by concentrating on the few memories of when he performed well, Harry may come to believe that he’s a good goal-keeper, even though the totality of the evidence that he was acquainted with suggests that he is not.

Now why would this strategy be thought to overcome the knowledge condition? The answer, I believe, would be assumed to lie with the thought-suppression element of the strategy. Thought-suppression may be conceived of as a *knowledge-subverting* strategy. For example, Harry shifting his attention off the memories of his bad performances is designed to undermine the knowledge that gets in the way of his having the welcome belief. How this happens, in the opinion of Van Leeuwen<sup>17</sup> and Whisner,<sup>18</sup> is that after perhaps repeated attempts at thought-avoidance Harry supposedly *forgets* about those performances.

The knowledge condition might still be considered as a stumbling block, however. If the thought-suppressor intentionally shifts her attention off a thought, doesn’t this mean that she will be intentionally not thinking about that thought? But granting the knowledge condition, this would imply that she’s aware of what she’s doing, i.e. not thinking about it.

---

<sup>16</sup> Lockie (2003: 131).

<sup>17</sup> Van Leeuwen (2008: 202).

<sup>18</sup> Wisner (1998: 196).

And paradoxically, this seems to imply that the thought is in mind, so that it has not really been forgotten after all.<sup>19</sup>

But there is little reason to entertain such worries. Here it may be useful to look at what thought-suppression typically involves. Daniel Wegner—one of the most prominent researchers on this issue—says that when people try to take their mind off something, they typically do so by putting it onto something else.<sup>20</sup> They turn their mind or attention to a ‘distracter’ which may absorb their attention. Now although this distraction seeking can be done intentionally, this doesn’t imply that the agent is intentionally not thinking about the unwelcome evidence when she succeeds in distracting herself from thinking about it. We can look on the action of suppressing a thought as analogous to the action of going to sleep. In both cases, we intentionally do things to *facilitate* certain results, i.e. certain thoughts being lost from consciousness, or our losing consciousness altogether. We may intentionally facilitate these results in the latter case by lying down in a dark, quiet room, and in the former, by shifting our attention onto something else. But the fact that we intentionally tried to suppress a thought doesn’t imply that when the thought is forgotten, we know that it is, or that we are intentionally not thinking about it, any more than our intentionally going to sleep implies that when we finally fall asleep, we know (or are aware) that we are asleep, or that we are intentionally sleeping.<sup>21</sup> These results are things that we intended without being things we are doing intentionally.

The ‘attentional account’ of self-deception, though popular, remains rather under-elaborated in the literature, and it remains to be seen whether it could represent a realistic strategy of self-deception. However, I believe it offers perhaps the best prospects for an intentionalist theory. The reason is that with the other strategies usually mentioned, nothing is done to subvert the knowledge associated with the deliberateness of the attempt, knowledge that would render the attempt futile. Suppressive strategies, however, are supposedly knowledge-subverting: they may be aimed at undermining knowledge of the considerations that support the unwelcome belief, perhaps by undermining memory of

---

<sup>19</sup> For similar worries, see Pugmire (1969: 346) and Reilly (1976: 393).

<sup>20</sup> Wegner (1994: 12 & 60).

<sup>21</sup> Suicide is another example. One commits suicide intentionally, but can’t know that one has succeeded when one has.

those considerations. Whether we have those abilities for mental manipulation, however, is debatable,<sup>22</sup> and it may ultimately be an issue for psychologists to have the last say on.<sup>23</sup>

**Kevin Lynch**

*Warwick University*

Kevinlynch405@eircom.net

## References

- Anscombe, G.E.M. (1966) *Intention*, Oxford: Basil Blackwell.
- Bach, K. (1981) 'An Analysis of Self-Deception', *Philosophy and Phenomenological Research*, 41, 351-370.
- Barnes, A. (1997) *Seeing Through Self-Deception*, Cambridge: Cambridge University Press.
- Bratman, M. (1984) 'Two Faces of Intention', *Philosophical Review*, 93, 375-405.
- Champlin, T.S. (1988) *Reflexive Paradoxes*, London: New York: Routledge.
- Champlin, T.S. (1977) 'Self-Deception: A Reflexive Dilemma', *Philosophy*, 52, 281-299.
- Donnellan, K.S. (1963) 'Knowing What I am Doing', *Journal of Philosophy*, 60(14), 401-409.
- Gustafson, D. (1975) 'The Range of Intentions', *Inquiry*, 18, 83-95.
- Hamlyn, D.W. (1971) 'Self-Deception', *Proceedings of the Aristotelian Society* (Supplementary volume 35), 45-60.

---

<sup>22</sup> Wegner's empirical studies of people's attempts to suppress unwanted thoughts (1994) have highlighted how difficult people find it to rid themselves of an unwanted thought or make themselves forget an unwanted memory.

<sup>23</sup> Thanks to Johannes Roessler, and to the audiences of the Second European Graduate School, Bochum Germany, and the Warwick philosophy department WIP seminar group, for useful comments on this paper. This paper was supported by a Warwick Postgraduate Research Scholarship.

- Hampshire, S. (1970/1959) *Thought and Action*, London: Chatto and Windus.
- Lockie, R. (2003) 'Depth-Psychology and Self-Deception', *Philosophical Psychology*, 16, 127-148.
- Mele, A.R. (1998) 'Two Paradoxes of Self-Deception', IN Jean-Pierre Dupuy (ed.) *Self-Deception and the Paradoxes of Rationality*, Stanford: CSLI Publishing, 37-58.
- Mele, A.R. (1997) 'Real Self-Deception', *Behavioral and Brain Sciences*, 20, 91-102.
- Mele, A.R. (1987) 'Recent Work on Self-Deception', *American Philosophical Quarterly*, 24, 1-17.
- Moran, R. (2001) *Authority and Estrangement*, Princeton; Oxford: Princeton University Press.
- Miller, A.R. (1980) 'Wanting, Intending, and Knowing What One is Doing', *Philosophy and Phenomenological Research*, 40, 334-343.
- Pugmire, D.R. (1969) "'Strong" Self-Deception', *Inquiry*, 17, 339-361.
- Reilly, R. (1976) 'Self-Deception: Resolving the Epistemic Paradox', *The Personalist*, 57, 391-394.
- Siegler, F.A. (1963) 'Self-Deception', *Australasian Journal of Philosophy*, 41, 29-43.
- Steffen, L.H. (1986) *Self-Deception and the Common Life*, New York etc.: Peter Lang.
- Van Leeuwen, D.S.N. (2008) 'Finite Rational Self-Deceivers', *Philosophical Studies*, 139, 191-208.
- Wegner, D.M. (1994) *White Bears and Other Unwanted Thoughts*, New York; London: The Guilford Press.
- Whisner, W. (1998) 'A Further Explanation and Defense of the New Model of Self-Deception: A Reply to Martin', *Philosophia*, 26, 195-206.

## SYNAESTHESIA AND THE RELEVANCE OF PHENOMENAL STRUCTURES IN PERCEPTION \*

Michael Sollberger

### Abstract

The aim of the present paper is to sketch a new structural version of the Representative Theory of Perception which is supported both by conceptual and empirical arguments. To this end, I will discuss, in a first step, the structural approach to representation and show how it can be applied to perceptual consciousness. This discussion will demonstrate that perceptual experiences possess representational as well as purely sensational properties. In a second step, the focus will switch to empirical cases of synaesthesia. In particular, I will stress that certain synaesthetic experiences are well-suited to corroborating a structural account of the perceptual mind. The overall picture that emerges in this paper prompts a new view of perceptual consciousness that is ruthlessly structural.

### 1. Introduction

Perceptual states seem to put us in direct contact with ontologically and causally independent empirical objects and their properties, such as, the shape of a table, the smell of a flower, the pitch of a sound, etc. That much seems uncontroversial. However, controversy arises as soon as one wonders how to conceive the metaphysics of the objects and properties we are aware of in conscious attentive perception. Ultimately, this controversy concerning the nature of perceptual consciousness derives from the arguments from illusion and hallucination, as well as from the causal argument.<sup>1</sup> In fact, what these *arguments from perceptual error* are supposed to highlight is that perception cannot be what it intuitively seems to be, namely, the direct awareness of objects in the external world that exist here and now. The arguments thus seek to establish that empirical objects fail to directly determine the perceptual consciousness of the perceiver. Instead, what we as perceivers are said to be immediately aware of are *inner mental items*, usually called sense-

---

\* The work on this paper has been supported by the Swiss National Science Foundation (SNSF), grant nr. 100011-117611. Thanks to Michael Esfeld, Gianfranco Soldati, an anonymous referee, and especially the attendees of the Second European Graduate School in Lausanne, Switzerland, for criticism and advice.

<sup>1</sup> See Smith (2002).

data, sensa, sensibilia, qualia, phantasms, impressions, ideas, or what have you. Against this background, the metaphysical status of sensuous properties becomes highly controversial.

Due to lack of space, I shall not go into further details here. In what follows, I will simply take for granted that external objects fail to have any *direct* bearing on perceptual consciousness (I have argued for this at length in Sollberger 2008). Typically, this assumption has been taken to lead to indirect realism and, more specifically, to the so-called *Representative Theory of Perception* (henceforth called RTP): a perceptual experience is an inner sensory experience of the perceiver S that has been appropriately caused by the external physical object x, and the phenomenal properties of which S is directly aware in attentive perception are properties of inner sensory experiences and not of objects experienced. That is, phenomenal properties are neither identifiable with nor reducible to the physical properties of objects experienced. Furthermore, the mental item or state of which S is directly aware is said to *represent* states in the external physical world. Sensory states are perceptual proxies that S immediately senses and by virtue of which S mediately perceives the physical world.<sup>2</sup>

Based on these assumptions, the goal of the present paper is to sketch a new version of RTP. More particularly, I want to make a case for a *structural* understanding of perceptual consciousness by dwelling on two main issues: a) the structural account of mental representation and b) empirical cases of synaesthesia. Hence, the paper is meant to shed light on the nature of the representation relation which RTP supposes holds between the inner phenomenal and the external physical realm.

Of course, some readers will disagree with the starting point of this paper and reject any form of RTP out of hand. I shall not attempt to convince them of the contrary.<sup>3</sup> Instead, those readers are invited to read the paper as dealing with the following conditional claim: if one admits the validity of RTP, then there are good reasons to consider perceptual consciousness in structural terms. In addition, much of what will be said is also relevant to

---

<sup>2</sup> For a representative survey of RTP, see, for instance, the papers in Wright (1993).

<sup>3</sup> RTP has recently gone out of fashion as a theory of perceptual consciousness. To my mind, its current pariah status is largely unjustified since the arguments in favour of RTP and the replies that have been provided to various objections to it in the past have been almost totally neglected and ignored by the philosophical community. For more on this, see Wright (2008).

perception and perceptual consciousness *per se* and not essentially tied to RTP's specific framework. Having cleared up these caveats, let us start by considering the topic of mental representation.

## 2. Mental representation

In order to understand the nature of perceptual states, one is well-advised to take into account current empirical data from the *cognitive sciences*. After all, perceptual states are complex information-carrying states that enable cognitive systems to successfully navigate through their environment. Therefore, an adequate philosophical analysis of perception should not ignore the context of cognition and cognitive explanations.

Importantly, cognitive explanations of behaviour routinely refer to *internal mental representations* and relevant operations over them. That is, cognitivists posit mental representations in order to explain the problem-solving behaviour of intelligent creatures. At bottom, 'a representation is something that stands in for and carries information about what it represents, enabling the system in which it occurs to use that information in directing its behaviour'.<sup>4</sup> Perceptual states are thus conceived of as mental states that represent the external world by means of internal representations.

Of course, there is an ongoing debate concerning the correct account of representations. Several theories have been proposed: causal, functional, teleofunctional, and structural theories.<sup>5</sup> To my mind, the most promising theory on the market is the *structural* account of representation, according to which representation is understood as a *transfer of structure*.<sup>6</sup> More precisely, the theory maintains that there must be a mapping (correspondence-function) from objects in the represented domain *B* to objects in the representing domain *A*, such that at least some relations in *B* are *structurally preserved* in *A*.<sup>7</sup> This mapping or correspondence-function from *B* to *A* can be conceived as a *homomorphism*, e.g., *A* is a

---

<sup>4</sup> Bechtel (2001: 334).

<sup>5</sup> See Fodor (1987), Cummins (1989), Dretske (1995), and Cummins (1996), respectively.

<sup>6</sup> Advocates of the structural account include, among others, Bartels (2005), Cummins (1996), Gallistel (1990), and Palmer (1978).

<sup>7</sup> A structure  $U = (O, R)$  is characterized by two elements: a non-empty set *O* of objects that constitute the domain of *U* and a non-empty set of relations *R* on *O*.

homomorphic image of  $B$ .<sup>8</sup> Maps are paradigmatic examples of structural representations: a city map of London can represent the streets and houses of London in virtue of preserving a spatial structure that is a homomorphic image of London. Likewise, a photo can represent its subject matter in virtue of mirroring its relevant structure.<sup>9</sup> In short, the idea is that  $A$  represents  $B$  *only if*  $A$  is a homomorphic image of  $B$ , with  $A$  and  $B$  being defined as structures.

More precisely, this means that the content of a representation is specified by an abstract structural description. This further implies that representational content is not primarily about particular *individuals*, but about *structures* and *relational properties*. Particular individuals are represented only derivatively, namely, in virtue of the fact that they occupy *corresponding logical spaces* in the structurally defined domains  $A$  and  $B$ . In fine, the present structural account prompts the conclusion that the relations in which objects stand take representational priority over the objects as such.

The structural account of representation needs further to distinguish between the *content* and the *target* of a representation (see especially Cummins (1996) for this issue). Without this distinction, the account remains incomplete, because an infinite number of external physical structures might in principle be homomorphic to a given content. In other words, a particular content underdetermines its target. This problem can be solved as follows: a given content determines all the *potential* targets of a representation, and additional *contextual* factors, such as, causation, intention, cognitive abilities of the organism, etc., fix the *actual* target of the content. Structural similarity or homomorphism on its own is therefore insufficient for representation; it must be supplemented by further contextual factors by means of which the actual target of the structurally defined content is unambiguously fixed.

---

<sup>8</sup> The concept of homomorphism derives from mathematics: in abstract algebra, a homomorphism is a mapping between two algebraic structures of the same type that preserves all the relevant structure; it maps identity elements to identity elements, and it is compatible with all binary operations. For a formal definition of homomorphisms and a detailed discussion about how it can be used for modelling the representation relation, see Bartels (2005).

<sup>9</sup> 'Relevance' is of course not a mathematical concept but has to be added as a further element in order to arrive at a substantive theory of representation. I shall come back to this in a minute.

However, nothing that has been said so far about the structural account of representation suffices to render a representation *distinctively sensory* or perceptual in character. I propose the following: what renders a representation distinctively perceptual is that it *provides guidance for action with regard to x*. That is, the representation must enable S to focus her activities on x; such as, perceptually tracking and demonstratively pointing at x. This inside-out perspective acknowledges the importance of action for a representation-consuming system.<sup>10</sup> Three conditions are thus required for a mental state *A* to *perceptually* represent the external physical world *B*:

- i) *A* must share relevant structural features with *B*
- ii) *A* must have been appropriately caused by *B*
- iii) *A* must provide guidance to S in taking action with regard to *B*

More specifically, this means that i) determines the content of a representation, ii) fixes the actual target of the representation and iii) is what makes the representation distinctively perceptual. Applied to RTP, this yields the following modified account: A subject S can *navigate* the external world because internal sensory experiences are informative by preserving biologically relevant structural properties of the external world, and these structural properties can be decoded and exploited by the representation-consuming system S in order to guide S's actions with respect to the external world.

Before applying this picture to perceptual consciousness, one point should be noted: in the present context of perception, it is the science of *neuroethology* that attempts to provide an answer to the question of which structural properties are biologically relevant to S.<sup>11</sup> This means that the concept of 'relevance', which the structuralist has to define in order to make clear which structures are preserved by perceptual representations, will be spelled out in empirical terms. It is not necessary for present purposes to deal with the intricate details of this empirical enterprise.<sup>12</sup> What matters is that the science of neuroethology can be

---

<sup>10</sup> My account must be distinguished from Anderson & Rosenberg's (2008) guidance theory of perception. In contrast to their theory, which claims that the content of a representation is determined by guidance for action, the present proposal implies that content is structurally determined and guidance for action enters the scene solely in order to explain what makes a representation distinctively perceptual.

<sup>11</sup> Keeley (2000).

<sup>12</sup> The interested reader may consult Keeley (2000) for the corresponding literature on neuroethology.

relied on by adherents of the structural approach to show that a clear definition of ‘relevant structural properties’ is available for the domain of perceptual representations.

### 3. Phenomenal content of perceptual states

In accordance with the requirements of the cognitive sciences, I shall thus take it for granted that perceptual states are *representational states*. This means that perceptions can represent the world veridically or falsidically. The structurally defined representational content of a perceptual experience is a condition of satisfaction of the experience: an experience is veridical iff the world satisfies the condition. That is, S’s experience of an x standing in relation R to y is veridical iff there is an x that stands in R to y. Let us further assume that it is highly plausible to apply this representational scheme to perceptual consciousness as well.<sup>13</sup> Then, the phenomenal character of an experience can determine a condition of satisfaction for the experience, and this condition of satisfaction is its *phenomenal content*.

With this assumption at hand, the structural framework laid out so far entails that the phenomenal character of an experience can determine a representational content only by means of its *structural* properties. This insight is key to a proper understanding of structural phenomenal content: phenomenal properties *per se* do not represent anything! Phenomenal properties like redness, roundness and so forth are nothing but the non-representational atomic building blocks of the representational structure – i.e. they are *non-epistemic raw feels*. Fundamental units or building blocks are required to make up the structure by instantiating numerous relational properties amongst themselves. This is what phenomenal properties do: they stand in multitudinous relations of similarity and difference to each other and thus build up the structure of the phenomenal character of an experience. Yet, it is only the phenomenal structure *qua structure* that is able to represent the empirical world.

Consider an example: S is phenomenally aware of a red apple placed on a round table in front of her. Redness, roundness, and several further phenomenal properties figure in S’s perceptual consciousness. The present structural account underscores that what matters for

---

<sup>13</sup> Siewert (1998: chapter 7).

representational purposes is not the particular ‘feel’ of phenomenal red. Rather, it is by means of *relational* facts – e.g., red is more similar to orange than to green, roundness is more similar to ovalness than to squareness etc. – that phenomenal character determines a representational content. Phenomenal properties exhibit similarity/difference relations amongst themselves and thereby instantiate relational properties that ground the structure of phenomenal character. Phenomenal properties are thus brute sensational units whose intrinsic properties, e.g., their particular feel or what-it-is-likeness, give rise to the representational nature of phenomenal character by building up a phenomenal structure. In sum, inner sensory experiences have both representational and non-representational properties.

Before going on, a short remark about the similarity/difference relations is in order. If asked ‘Why should the phenomenal properties be similar specifically in these respects’, the adherent of the structural framework cannot make reference to external physical objects. That is, the explanation that phenomenal properties are similar to each other because they supposedly represent things and properties that stand in the relation of similarity and difference to each other is unavailable to her. Instead, the structuralist must either a) bite the bullet and treat this as a primitive fact about phenomenal properties or b) speculate that a future theory about the mind/brain relation may come up with such an explanation. Both options have their price, to be sure, but they nevertheless present intelligible positions the structuralist can consistently endorse.

Let’s now summarise what has been said so far. We then arrive at the following definition of *veridicality*:

S’s perceptual experience of the  $\psi$ -type is veridical *iff* there exists a homomorphic mapping function from the structure instantiated by the external physical world to the structure instantiated by the phenomenal character of S’s experience, and the experience has been appropriately caused by the external physical structure that usually causes experiences of the  $\psi$ -type in S.

It is noteworthy that veridicality thus understood has both conventional and revisionary aspects. Like *conventional* accounts, the above definition requires a match between the content of an inner sensory state and properties of the external world, and this content

match must causally depend upon those worldly properties in the right way. Up to this point, the structurally construed notion of veridicality is still in line with tradition.

Much more controversial, however, are its *revisionary* aspects. Intuitively, doing justice to the phenomenology of experiences seems to imply that what is conveyed to us as perceivers in perceptual experience is a) that the world contains individual objects that instantiate intrinsic properties and b) that these individual objects are the primary focus of perception. Yet, contrary to what is stressed in a), the structural conception of veridicality yields that objects are stripped of their intrinsic properties, i.e., objects determine accuracy conditions only by means of the relations they enter into and not by virtue of their intrinsic properties. This means that intuitions about content that rely solely on phenomenology are misleading and have to be revised. Furthermore, as regards b), one can see that the structural account reverses the order of *experiential salience* involving individual objects and relational properties since it implies that the representational focus is primarily on relational properties and, as previously shown, merely secondarily or derivatively on individual objects. This highlights again that we cannot trust our intuitions about phenomenal content without certain reservations.

Hence, the prize to pay is *partial phenomenal inadequacy*, since structural content does not do justice to the phenomenology in all respects. But some might wonder why one should bite this bullet at all and accept such a revisionary account. Here is one such reason: It delivers the right answer to the empirically pervasive phenomenon of *shifted qualia*.<sup>14</sup> The structural account implies that the perceptual experiences of perceivers who are 'normal' in behavioural, biological and functional respects, but whose phenomenal properties have shifted by comparison, are really on the same *epistemic footing*. It has been rightly argued that there is no reason to *epistemically* privilege one group of perceivers over another group simply because their phenomenal properties are found to have shifted by comparison. The present account naturally accommodates this idea, for experiences of different perceivers which have shifted with regard to their qualia can equally well satisfy

---

<sup>14</sup> Hardin (2008).

or fail to satisfy the above definition of veridicality as long as they have a common structure.<sup>15</sup>

A further reason why one might prefer a structural account is that it is suggested by some empirical cases of synaesthesia. In the remainder of this paper, I would like to discuss this particular special case.

#### 4. Synaesthesia and the relevance of phenomenal structures

Briefly, synaesthesia is an intrinsically perceptual phenomenon where ‘stimulation of one sensory modality automatically triggers a perception in a second modality, in the absence of any direct stimulation to this second modality’.<sup>16</sup> Some phenomenal properties are *reliably and systematically* elicited in response to certain stimuli that are not elicited in non-synaesthetes. Synaesthetes can *hear colours, taste shapes, smell sounds*, etc. In principle, any pairing of the senses is possible, although coloured hearing, e.g., the pairing of sound and sight, is the most common combination. Consider subject MW: in addition to gustatory and olfactory properties, MW perceives tactile properties of weight, shape, texture, and temperature whenever he tastes or smells food. In MW, these sensory dimensions of touch experiences are *functionally related* to flavours and odours. For instance, he synaesthetically perceives the taste of spearmint as a ‘cool, glass column’, and lemon is like ‘a pointed shape, pressed into my hands. It’s like laying my hands on a bed of nails’.<sup>17</sup> Among other things, I want to argue that MW’s case can provide empirical evidence for the possibility of *cross-modal exchange of sensory properties* without the experiences becoming falsidical. Notice that this idea is more radical than the aforementioned case of shifted qualia, for it holds that sensory properties are not constitutively but only contingently associated with their respective sense modalities.<sup>18</sup>

---

<sup>15</sup> Note that a similar reasoning can be applied, *mutatis mutandis*, to hypothetical cases of *spectrum inversion*. See Palmer (1999) for more on the topic of inverted spectra.

<sup>16</sup> Baron-Cohen and Harrison (1997: 3).

<sup>17</sup> Cytowic (2002: 1).

<sup>18</sup> This idea is of course highly relevant for the question of how to individuate the senses. Space precludes a more detailed treatment of this topic.

To begin with, I want to stress that *some* synaesthetic experiences can be treated as veridical.<sup>19</sup> Synaesthetic experiences are a normal variant of human perception and ‘abnormal’ only in that they are statistically rare. Three reasons can be invoked: Firstly, one has to take seriously *subjective reports* of synaesthetes. After all, some synaesthetes have an unshakable conviction that what they synaesthetically perceive is real and valid, and not hallucinatory or illusory. Neither the phenomenology nor the content of these synaesthetic experiences indicate to the subject that something weird or outlandish would be occurring. In short, there is nothing special about synaesthetic experiences that would prompt synaesthetes to treat them differently from non-synaesthetic perceptual experiences. What is more, synaesthetes have been extensively studied by empirical researchers in recent years. These results clearly indicate that there is so far no scientific reason to doubt their subjective reports.

Secondly, it is often true that synaesthesia *enhances* several cognitive capacities of its bearer: the additional synaesthetic sense enhances the ability of reading, writing and spelling, and it also expands the memory faculties by acting as a mnemonic device.<sup>20</sup> This seems to suggest that synaesthesia is certainly not a maladaptive biological trait. Quite the opposite, it can mean an *adaptive advantage* for its bearer.<sup>21</sup> One further reason, then, why one should not treat such synaesthetic experiences as falsidical.

Thirdly, it is instructive to approach synaesthesia in terms of *evolution* and *natural selection*. From a purely evolutionary perspective, the goal of perception is to *maximize fitness*, i.e., raising more offspring! Perception must be viewed as a niche- and problem-specific cognitive function whose purpose is to enhance fitness.<sup>22</sup> Importantly, S is able to survive and reproduce only if S can *successfully interact* with the world. And successful

---

<sup>19</sup> A caveat: the following, admittedly sketchy, description of synaesthetic experiences and of MW cannot be generalized to cover all cases of synaesthesia. It refers only to those synaesthetes who attribute the synaesthetic component of their experience to the *distal* object itself and who do not take their synaesthetic experiences to be illusory or hallucinatory (see especially Cytowic 2002: chapter 2). The phenomenology of synaesthetes is heterogenous, highly idiosyncratic and difficult to describe adequately. In this sense, then, keep in mind that my proposal is *one* way one might interpret subjective reports given by *some* synaesthetes.

<sup>20</sup> Cytowic (2002: 29).

<sup>21</sup> Recent work has strongly suggested that synaesthesia is an inherited condition. The potentially beneficial trait can thus be carried over from parents to offspring. See Harrison and Baron-Cohen (1997).

<sup>22</sup> (cf. Hoffman 2009).

interaction is possible only if the subject can adequately *discriminate* between objects and properties. For example, based on perceptual information, S can see, reach, grasp, and finally eat the red apple in front of her. This discriminatory behaviour is an instance of successful interaction with the world based on which S is, in the long run, able to survive and reproduce.

It is crucial to note that some synaesthetes can, up to a certain extent, perform the same discriminatory tasks as non-synaesthetes, based on their synaesthetically induced phenomenal properties. Consider subject MW. As a matter of fact, MW likes cooking. But the way he cooks is quite intriguing for he prepares food according to the *shape* of the food and not its flavour. By trial and error, he administers different seasoning in order to change the shape of, say, the chicken, for instance, making it rounder, sharpening corners in order to apply more heft to the vertical component, or adding some points to the overall shape.<sup>23</sup> This ‘cooking-according-to-shapes’ is impressive, for it highlights that MW’s *tactile* synaesthesias allow him to execute the same activities non-synaesthetes perform with the help of olfactory and gustatory properties. The synaesthetically evoked tactile phenomenal properties guide MW in taking the same actions with regard to food as ‘normal’ perceivers based on gustatory and olfactory properties. Hence, with regard to the coarse-grained behavioural context of cooking, synaesthete MW and any other non-synaesthete can be *functionally equivalent!*

Indeed, MW displays a discriminatory behaviour with regard to food that is an instance of successful interaction with the world. As such it contributes to MW’s survival and reproduction and can thus be treated as a fitness-enhancing perceptual capacity. Finally, that’s why, from an evolutionary point of view, MW’s synaesthesia is *on a par* with non-synaesthetic experiences! And given that we unhesitatingly treat most everyday non-synaesthetic experiences as veridical, it follows that the evolutionary perspective provides reasons for treating MW’s synaesthetic experiences as veridical as well.

In sum, the aforementioned three reasons represent *cumulative* justification for regarding MW’s synaesthetic experiences as accurate perceptual experiences. If this is accepted, in virtue of what feature can both ‘normal’ experiences and MW’s synaesthetic

---

<sup>23</sup> Cytowic (2002: 86).

experiences be veridical? After all, the sensory properties associated with synaesthetic and non-synaesthetic experiences are phenomenally quite distinct from each other. It seems obvious that the only relevant *experiential* feature they do have in common is phenomenal structure. The answer is: It is reasonable to claim that the structure of perceptual consciousness is rendered manifest in discrimination tasks.<sup>24</sup> In our example, MW is able to correctly season the chicken in virtue of the fact that MW can mentally point to the chicken. The phenomenal character of MW's experience instantiates relational properties that enable MW to demonstratively tag the chicken and thus discriminate it from its background. Consequently, if two subjects can be functionally equivalent within a certain behavioural scope, as it is the case for MW and non-synaesthetes, this is evidence for the fact that their experiences instantiate type-identical phenomenal structures. Therefore, synaesthetic and non-synaesthetic experiences can instantiate type-identical phenomenal structures.

The same conclusion can be reached more straightforwardly by acknowledging that there is a reliable and systematic functional relationship between flavours/odors and the synaesthetically induced tactile properties in MW. The existence of such cross-modal functional correspondences is sufficient to show that taste and smell are mapped onto touch. Hence, MW's synaesthetic tactile experiences can in principle build up structures of the same abstractly described type as his non-synaesthetic olfactory and gustatory experiences.

It follows that the only experiential feature MW's synaesthetic touch experience shares with the non-synaesthetic taste and smell experience of 'normal' perceivers is phenomenal structure. As a result, structure turns out to be the only feature that really matters with regard to the veridicality of perceptually conscious states. This, then, is how adherents of RTP may draw upon MW's empirical case of synaesthesia in order to back up the structuralist account of perceptual consciousness.

I have presented my argument by specifically dwelling on MW's case, but it is clear that the scope of MW's functional equivalence with non-synaesthetes is quite restricted.

---

<sup>24</sup> Notice that such discriminatory tasks are used in psychophysics in order to establish so-called psychophysical maps for individual subjects.

However, there is no reason why one should not consider cases beyond MW's limited framework of food and cooking. In so doing, certain cases of synaesthesia become suggestive and supportive of the possibility of what we might call *super-synaesthesia*: i.e., as relates to a synaesthete whose functional equivalence is not restricted to any particular range of behavioural context. That is, *super-synaesthetes* are conceived as having synaesthetic experiences that enable them to carry out the same range of successful interaction with the world as non-synaesthetic perceptual experiences. One might go further and stipulate that super-synaesthetes lost their non-synaesthetic experiences due to some brain damage, so that the super-synaesthete is only conscious of synaesthetically induced sensory properties. For example, if MW were such a super-synaesthete, he would only have tactile experiences whilst tasting and smelling physical things, and these tactile experiences would allow him, without restriction, to engage in exactly the same actions as regards the physical world as do non-synaesthetes based on their taste and smell experiences. The issue to be emphasized here is that, according to the structural account and the definition of veridicality presented above, super-synaesthetic experiences can count as truly veridical, although their sensory properties are *cross-modally exchanged* relative to 'normal' experiences.

Finally, I have tried to show that it is irrelevant how the external physical world is phenomenally represented, as long as the modelling is structure-preserving. Whether a given physical structure tastes like spearmint or tactually feels like a cool glass column or anything else is of no *representational* significance as long as the phenomenal character of the experience enables the subject to make the relevant discriminations between physical objects and their properties. Accordingly, one and the same physical stimulus may causally give rise to sensory experiences with wildly distinct phenomenal characters, and all of these experiences can be veridical. This is so because phenomenal properties *per se* are representationally inert – they are non-epistemic raw feels. What counts is that the

phenomenal character of the experience – be it tactile, auditory, olfactory, gustatory, visual, etc. – mirrors relevant structural properties of the external physical world.<sup>25</sup>

## 5. Conclusion

To sum up, according to the structural version of the *Representative Theory of Perception* I have sketched in this paper, inner sensory experiences of which S is directly aware in attentive perception represent the outside physical world by virtue of being *structurally similar* to it. By combining the structural account of mental representation with empirical cases of synaesthesia, I hope to have demonstrated how important an understanding of structure is to the theory of perception and perceptual consciousness.

**Michael Sollberger**

*Université de Lausanne*

michael.sollberger.2@unil.ch

## References

- Anderson, M. L. and Rosenberg, G. (2008) ‘Content and Action: The Guidance Theory of Representation’, *The Journal of Mind and Behavior* 29 (1/2), 55-86.
- Baron-Cohen, S. And Harrison, J. E. (eds.) (1997) *Synaesthesia: Classic and Contemporary Readings*, Oxford: Blackwell Publishers.
- Bartels, A. (2005) *Strukturelle Repräsentation*, Paderborn: Mentis.
- Bechtel, W. (2001) ‘Representations: From Neural Systems to Cognitive Systems’ IN W. Bechtel, P. Mandik, J. Mundale and R. S. Stufflebeam (eds.), *Philosophy and the Neurosciences: A Reader*, Oxford: Blackwell, pp. 332-348.
- Cummins, R. (1989) *Meaning and mental representation*, Cambridge, Mass.: MIT Press.

---

<sup>25</sup> It goes without saying that several objections would have to be addressed in order to defend this argument about synaesthesia more thoroughly. Due to lack of space, I must postpone this task for another occasion. My more moderate aim in this paper was just to outline the fundamental idea of the argument.

- . (1996) *Representations, Targets, and Attitudes*, Cambridge, Mass.: MIT Press.
- Cytowic, R. E. (2002) *Synesthesia: A Union of the Senses*, Cambridge, Mass.: MIT Press.
- Dretske, F. (1995) *Naturalizing the Mind*, Cambridge, Mass.: MIT Press.
- Fodor, J. A. (1987) *Psychosemantics*, Cambridge, Mass.: MIT Press.
- Gallistel, Ch. R. (1990) *The organization of learning*, Cambridge, Mass.: MIT Press.
- Hardin, C.L. (2008) 'Color Qualities and the Physical World', IN E. Wright (ed.), *The Case for Qualia*, Cambridge, Mass.: MIT Press, pp. 143-154.
- Harrison, J. E. and Baron-Cohen, S. (1997) 'Synaesthesia: a Review of Psychological Theories', IN S. Baron-Cohen and J. E. Harrison (eds.), *Synaesthesia: Classic and Contemporary Readings*, Oxford: Blackwell Publishers, pp. 109-122.
- Hoffman, D. D. (2009) 'The interface theory of perception: Natural selection drives true perception to swift extinction', IN S. Dickinson, M. Tarr, A. Leonardis and B. Schiele (eds.), *Object categorization: Computer and human vision perspectives*, Cambridge, NY: Cambridge University Press, pp. 148-166.
- Keeley, B. L. (2000) 'Neuroethology and the philosophy of cognitive science', *Philosophy of Science* 67 (Proceedings), 404-417.
- Palmer, S. E. (1978) 'Fundamental Aspects of Cognitive Representation', IN E. Rosch and B. B. Lloyd (eds.), *Cognition and Categorization*, New York: John Wiley & Sons, pp. 259-303.
- . (1999) 'Color, Consciousness, and the Isomorphism Constraint', *Behavioral and Brain Sciences* 22, 923-943.
- Siewert, Ch. (1998) *The Significance of Consciousness*, Princeton: Princeton University Press.
- Smith, A.D. (2002) *The Problem of Perception*, Cambridge, Mass.: Harvard University Press.
- Sollberger, M. (2008) 'Naïve Realism and the Problem of Causation', *Disputatio* 3 (25), 1-19.
- Wright, E. (ed.) (1993) *New Representationalisms: Essays in the Philosophy of Perception*, Aldershot: Ashgate.
- . (ed.) (2008) *The Case for Qualia*, Cambridge, Mass.: MIT Press.