2016, 2 (1)

ARGUMENTA

The Journal of the Italian Society for Analytic Philosophy

First published 2016 by University of Sassari

Produced and designed for digital publication by the *Argumenta* Staff

Argumenta is the official journal of the Italian Society for Analytic Philosophy (SIFA). It was founded in 2014 in response to a common demand for the creation of an Italian journal explicitly devoted to the publication of high quality research in analytic philosophy. From the beginning *Argumenta* was conceived as an international journal, and has benefitted from the cooperation of some of the most distinguished Italian and non-Italian scholars in all areas of analytic philosophy.

Contents

Editorial

In opening the second volume of *Argumenta* I am particularly pleased to report that the number of submissions to the journal is slowly but constantly increasing. I take this as the highest reward for all the work that the Editorial Board—especially the editorial assistants—and the colleagues who have generously acted as referees have been doing since the launch of the journal.

This issue includes the papers that over the last six months have been sent to the journal by scholars working in the analytic tradition all over the world, together with the papers sent by distinguished philosophers, who wanted thereby to express their personal acknowledgement of the official organ of the *Italian Society for Analytic Philosophy*. I heartily thank them all. All the papers have passed the double blind refereeing process.

In the editorial written for the previous issue of *Argumenta* I focused, among other things, on Hilary Putnam's death, which occurred last 13th March. In such cases, it is customary practice for journals to publish obituaries of a certain length, highlighting the importance of a given author for a specific field of study. In this issue the Editorial Board decided to do things slightly differently. Thanks to Gary Ebbs—one of the most acute exegetes of Putnam's thought—we are publishing a brief memorial of the great Harvard philosopher concentrating on one central element: his original view of the methods of inquiry in science and everyday life.

As usual, all the articles appearing in *Argumenta* are freely accessible and freely downloadable.

Buona lettura!

Massimo Dell'Utri
Editor

Argumenta 2, 1 (2016): 3
ISSN 2465-2334

Whose Existence? A Deflationist Compromise to the Fregean/Neo-Meinongian Divide

Giuliano Bacigalupo

University of Geneva

Abstract

The dispute between the Fregean and the Neo-Meinongian approach to existence has become entrenched: it seems that nothing but intuitions may be relied upon to decide the issue. And since contemporary analytic philosophers clearly are inclined towards the intuitions that support Frege's approach, it looks as if Fregeanism has won the day. In this paper, however, I try to develop a compromise solution. This compromise consists in abandoning the assumption shared by both Fregeanism and Neo-Meinongianism, namely that the notion of existence adds something to the content of a statement. To the contrary, we should think of existence as a redundant notion. In other words, I will argue that we should be deflationist about existence. Moreover, the kind of deflationism I propose relies on what I call the existence equivalence schema, a schema which follows the blueprint of the well-known truth equivalence schema. From such a perspective, we can say that Fregean philosophers rightly deny the status of a discriminating property to existence; and, conversely, Neo-Meinongians, too, rightly reject the view that existence is captured by quantification or expresses a universal property of objects. Finally, the argument that we should take a deflationist approach to existence builds upon an analysis of natural language (general) existential statements and their intuitive entailment-relations.

Keywords: Existence, Frege, Meinong, Deflationism

1. Introduction

There are few problems in philosophy on the solution of which there seems to be an overwhelming consensus. One of these exceptions is the interpretation of the notion of existence:[1] within contemporary (one should perhaps add analytic)

[1] In their statistical survey about what philosophers believe, Bourget & Chalmers (2014) draw attention to other, sometimes surprising, exceptions. Regrettably, however, they have not included in their survey the problem of existence. But they do record a consen-

Argumenta 2, 1 (2016): 5-24
ISSN 2465-2334

DOI 10.14275/2465-2334/20163.BAC

philosophy, almost everyone seems to follow Frege's approach. Existence, according to this view, is essentially captured by particular, so called 'existential' quantification. Or, in other words, existence expresses the second-order property of a concept, namely that it has at least one instance. Or, again in different words, existence expresses a universal property, co-extensive with the property of being self-identical.[2] As Quine remarked, the appropriate answer to the question as to what exists should be adamant: everything.

This does not mean that there is no alternative to the Fregean approach: the consensus is overwhelming but not universal. What I have in mind is the infamous Meinongian or, more precisely, Neo-Meinongian view of existence. According to authors who follow this heterodox tradition, such as Routley (1980), Parsons (1980), and Jacquette (1996), existence should be deemed an almost ordinary property of objects: the set of objects may be divided in two classes, those that exist and those that do not.[3]

The problem with both approaches is that neither has an argument to back up its crucial presupposition. To wit, they have no argument to explain why 'something is such and such' should be equivalent to 'there is something such and such' as Fregean philosophers maintain (see McGinn 2001: 21, Mendelsohn 2005: 113);[4] Or why statements of the form 'such and such a thing exists' should be considered on a par with statements of the form 'such and such a thing is red', as Neo-Meinongians claim (van Inwagen 2008: 58). Moreover, neither Fregean nor Neo-Meinongian philosophers have an argument to refute the opposite approach. For a while, it was believed that Fregean philosophers had the upper hand since Russell (1905) and Quine (1948) presented compelling arguments against the discriminating property-view of existence. But Neo-Meinongian philosophers (first of all, Routley 1980 and Parsons 1980) were able to counter-strike, so that we are back to square one. Thus, it seems that the only way we have to take a stance on this issue is by relying on our intuitions—or, at least, some philosophers involved in the debate saw themselves forced to draw this conclusion (see Lewis 1990: 27-28; Perszyk 1993: 178; van Inwagen 2008: 54).[5] And since there is a consistent majority of philosophers whose intuitions speak in favor of the Fregean approach, this account should be the preferred

sus with respect to the truth of classical logic, which, arguably, embeds a given view of existence, i.e. Frege's.

[2] One may want to draw lines between these formulations, so that they are not equivalent. For instance, one may want to stress how the second, but not the first and third one, commits us to the existence of concepts (see Branquinho 2012). In the present context, I put such considerations aside.

[3] For the sake of simplicity, I am focusing on the kind of Neo-Meinongianism that relies on a distinction between nuclear and extra-nuclear properties (i.e., the one defended by Routley, Parsons and Jacquette). The line of reasoning of this paper, however, may also be easily applied to other approaches that interpret existence as a discriminating property of objects, such as Zalta (1988) and Priest (2005).

[4] Throughout the paper, I am assuming that the expression 'there is' is existentially loaded. This is, for instance, rejected by Parsons (1980). To him, the unwarranted presupposition of Fregeanism should be formulated as follows: statements of the form 'there is something such and such' are equivalent to statements of the form 'there is something such and such which exists'.

[5] To say that the rival theory is unintelligible, as Lycan (1979: 290) and Horgan (2007: 620) consider it to be the case with Neo-Meinongianism (see Priest 2008a), may also be seen as pointing towards an irreducible clash of intuitions.

one. But there is no need to point out how this is a very unsatisfactory way of settling a philosophical dispute.

In what follows, I will attempt to break this stalemate. More precisely, I will try to argue for a compromise solution, or at least something that may be seen as a compromise: the notion of existence is neither a universal property nor a discriminating property of objects. The reason for this is that we should be deflationist about existence and abandon the assumption that we may find out something like the nature of existence. To be more precise, what I am going to argue for is a version of deflationism which relies on what I label the *existence equivalence schema* and its negative counterpart, the *non-existence equivalence schema*. As with the deflationary approach to truth (and falsity), whose blueprint I am following, the equivalence schemata provide us with everything we can and should say about existence.

The structure of the paper is as follows. First, I develop an argument that relies upon an analysis of quantified, i.e., general statements to support the deflationist approach to existence. Second, I show how the same analysis may be applied to different sets of statements, and, most crucially, to a set of statements which involve modal notions. Third, I address the challenge raised by intentional statements about indeterminate objects—a challenge which is notably exploited by Neo-Meinongians in order to further their stance. Finally, I underline the difference between the version of deflationism defended in the present paper and the one recently advocated by Thomasson (2014).

As the reader will have noticed, what is conspicuously missing from the picture are singular statements, i.e., statements which are allegedly about definite objects. The reason is that the author of this paper is an acolyte of a different heresy than the Neo-Meinongian one, namely descriptivism: I do not believe that there are such things as genuine singular statements; these are just hidden quantified statements.[6] If you like, you may thus think of this paper as proving—if anything—something about the notion of general existence. For any alleged notion of singular existence, a different account would have to be provided.

2. A Raw Intuition

Let me start with a terminological remark. Throughout this paper, by 'existential statement' I am referring to a statement in which the verb 'to exist' is embedded. Hence, my characterization is a strictly linguistic one. This I take to be the most suitable definition of existential statements since every statement which does not wear, so to speak, its existential character on its sleeve may be cast as a statement with the verb 'to exist'. For instance, a statement such as 'an existing dog is on the street' is meaning-equivalent with the statement 'a dog on the street exists'. Or the statement 'there are some dogs', where existence seems to be expressed by the expression 'there are', is also meaning-equivalent to the statement 'some dogs exist'. Finally, the same applies to statements in which the noun 'existence' is embedded: the statement 'the existence of dogs is uncontroversial' may be rephrased as 'it is uncontroversial that dogs exist'. Notice, moreover, that I am consciously leaving statements with a particular quantification out of this list: I am refraining from saying that 'something is x' is meaning-equivalent to 'there are xs' or 'some xs exist'. Indeed, this is what Fregeanism

[6] For a recent defense of descriptivism, see Orilia (2010).

and Neo-Meinongianism are arguing about, so that we cannot and should not take it for granted. The other equivalences, on the other hand, I take to be theoretically neutral.

Now, even before getting started on any philosophical lucubration about existential statements, I assume we would all agree that there is something peculiar about existential statements. And, in fact, there have been many attempts at rationalizing this difference well before Frege and Meinong or the Neo-Meinongian philosophers. To give just a few illustrious names, David Hume, Immanuel Kant and Franz Brentano have all tried to develop a philosophical theory which accounts for this peculiarity. The question with which I would like to start is thus the following: how can we give more substance to this shared intuition about the peculiarity of existential statements without bringing into play a given philosophical theory about them? Can we really do nothing better than say that this is our intuition? In other words, what I am looking for is some pretheoretical intuition about the difference in the behavior of existential statements and non-existential ones, which may be the reason for our shared intuition about the peculiarity of the former—or, if not *the* reason, at least *a* reason.

First, let us consider the two following existential statements:

(1) Something red does not exist.
(2) It is not the case that something red exists.

It seems to me that there is a very strong connection between these two statements: if one of them is true, the same holds for the other, and conversely, if one of them is false, so is the other. At least *prima facie*, (1) and (2) mutually imply one another. Or, in more technical terms, the internal negation seems to be interchangeable with the external one. Let me provide a few more standard examples to substantiate this claim. To say that something which is golden and a mountain does not exist (shorter: a golden mountain does not exist) is equivalent to saying that it is not the case that a golden mountain exists. To say that something which is round and square does not exist (shorter: a round square does not exist) is equivalent to saying that it is not the case that a round square exists (and so on and so forth).

Someone may challenge the reading just advanced since it seems easy to point to a situation in which both (1) is true and (2) false. As it happens, the actual state of things seems to be just the right one to accomplish this feat. We would all agree that, on the one hand, red dragons do not exist and, on the other hand, some red things, such as for instance traffic lights, do. Thus, why not assume that (1) is true because red dragons do not exist, while (2) is false because a red traffic light exists? However, I take this to be a misconstrual of (1). If someone were talking about red dragons, he would have to make it explicit. Thus, he should not say 'something red does not exist' but rather 'something which is a red dragon does not exist'. But this would be a very different claim than affirming (1).

The same point may be clarified in the following way. While having coffee with a friend of mine, we stumble upon the topic of redness, upon which I claim 'something red does not exist!' My friend then objects to my claim by pointing to the traffic light in front of the coffee shop. But then I go on to say that, of course, red traffic lights exist, but red dragons do not—which should be enough to make my claim warranted. What would my friend's reaction to this explanation be? I suppose he would give me an incredulous stare and say something

along the following lines: 'OK, what you really meant was that a red dragon does not exist. Yet this is a rather irrelevant remark to the topic of redness which we were just starting to discuss'.

Let us now turn to a very similar pair of sentences which, however, are not existential:

(3) Something red is not round.

(4) It is not the case that something red is round.

I gladly concede that this second pair, as probably the first, too, may sound awkward. Below, I provide a possible coffee shop scenario in which they may be uttered. Now, the difference between this second couple of statements and the first one is striking. Notwithstanding the very similar structure, we have lost the mutual—and for that matter any kind of—implication: the truth of (3) is compatible with both the truth and the falsity of (4). Conversely, the truth of (4) is compatible with both the truth and the falsity of (3). For it could always be the case that nothing is red. In other words, internal and external negations are no longer interchangeable.

True, someone may be tempted to interpret (3) as a hidden hypothetical, namely as really meaning that if something is red then it is not round. It may then be argued that such a hypothetical would indeed imply (4). Please set this interpretation aside: (3) should be read literally. An example of a literal reading of (3) and (4) is the following. I am still sitting in the same coffee shop as before and my friend points out that something red is round, namely the red traffic light. Without having any intention to contradict him, but just for the sake of conversation, I then say that it is also true that something red is not round, namely the red sports-car parked in the second row. This is the pre-theoretical linguistic intuition about the peculiarity of existential statements with which I wish to start my discussion.

I would like to stress that I am very well aware that not everyone would share this intuition. This is especially the case with philosophers, whose intuitions about existential statements have already been thwarted in one direction or another by their own theory about existence. Moreover, philosophers may stress that it is only on the background of a theory about existence that we may test the mutual implication of (1) and (2). For these reasons, I am labeling the intuition in question as *the raw intuition*. Now, I assume that even those who reject the raw intuition or have qualms about it should be interested in why one *may* have such an intuition. Thus, I will ask them to indulge me for a little while. I will come back to their worries at the end of section 2.3.

2.1. Fregeanism

Having introduced you to the raw intuition, I would like to explore how a Fregean and a Neo-Meinongian philosopher might make sense of it. This, moreover, will provide us with the opportunity to rehearse some theses and arguments of these two arch-enemies.

Let us start with Fregeanism. According to this approach, we should go Procrustean and amputate (1) from our language. As classical logic (the formal arm of Fregeanism) teaches us, there is only room for a universal predicate of existence in our formal language, which may be defined by means of quantification and identity (Hintikka 1966):

$E!(a) =_{def} \exists x(x = a)$

Thus, the formalization of (1), if we take the verb 'to exist' as expressing the predicate of existence, would yield us a contradictory statement (see, e.g., Lewis 1990: 25), which should be formalized as follows:

(1*) $\exists x(Rx \wedge \sim E!x)$

On the other hand, (2) may be formalized without further ado into the somewhat redundant (2*):

(2*) $\sim \exists x(Rx \wedge E!x)$

Thus, according to Fregeanism, the puzzle of the raw intuition is resolved to the extent that, while stating (1), we cannot really mean what we say. Rather, if we are reasonable agents, what we mean is (2). Or, from a different perspective, one may prefer to say that language is misleading because both (1) and (2) express the same logical form (2*).[7]

What is the philosophical reason for the Fregean approach to the raw intuition, namely that we should get rid of (1)? As it happens, it is nothing over and above a strong intuition which philosophers and non-philosophers alike seem to share, namely the *predication principle* (PP):[8]

(PP) If something instantiates a property, then it exists.

Arguably, this is the principle Russell (1919: 170) had in mind while talking of a robust sense of reality. A most common formulation of the principle is to say that statements of the form 'something is such and such' are equivalent to statements of the form 'there is such and such a thing' (see Frege 1883?: 63).

2.2. Neo-Meinongianism

Let us turn to a short exposition of the Neo-Meinongian strategy. Neo-Meinongianism rejects (PP): non-existent objects may instantiate properties. Thus, the domain of objects is divided into two classes, namely, existent and non-existent ones. As a consequence, quantification is existentially neutral in that it has to range over all objects. To a Neo-Meinongian such as Routley, (1) does not have to be interpreted away or even be amputated from our language. Rather, (1) finds a streamlined logical interpretation as (1**) (read Px as the existentially neutral particular quantifier):

(1**) $Px(Rx \wedge \sim E!x)$

The fact that Neo-Meinongianism is in a position to provide such a streamlined logical interpretation of (1) does indeed count as one of the main advantages of this position, as stressed by Routley (1980: 31-32). The logical form of (2), on the other hand, turns out as follows:

(2**) $\sim Px(Rx \wedge E!x)$

[7] McLeod (2011: 260) rightly stresses this second option: a Fregean philosopher does not have to claim that the expression 'some' in natural language always must have existential import (although Frege himself did). However, the difference between these two strategies is minimal. If pressed, a Fregean may only provide the following answer to the question as to why the superficial grammatical form (1) should be seen as hiding the deep logical structure (2*), namely that otherwise it would have to be interpreted as the contradictory (1*).

[8] Routley (1980: 21) labels (PP) as the Ontological Assumption, thus introducing some terminological bias against it.

Notice that a Fregean philosopher cannot refute a Neo-Meinongian philosopher by recurring to (PP). The problem is that Neo-Meinongians reject (PP), so that this strategy would be question-begging.[9] From our perspective, instead, the problem a Neo-Meinongian is confronted with is the following: he has to show how some additional premises are responsible for the fact that from (1) we may infer (2), and the other way around—premises which should of course not allow to infer (4) from (3) and (3) from (4).

One such premise is the *restricted characterization principle* (RCP):[10]

(RCP) For any condition α that does not embed extra-nuclear properties, an object satisfies exactly this condition.

I do not wish here to enter into the details as to why (RCP) must be restricted to nuclear properties and strengthened so that the object instantiates no other property besides those embedded in the characterization (the object satisfies the condition *exactly*).[11] Nor am I interested in the philosophical reason behind (RCP), namely that, against (PP), for any nuclear property, an object exemplifies these properties. Here, I would simply like to point out how (RCP) is needed in order to make sense of the raw intuition. Indeed, it is only if we concede (RCP), or something sufficiently close to it, that we are in a position to say that, if a given condition α is not satisfied by an existing thing, then it is satisfied by a non-existing one. And this is exactly what we need in order to infer (1) from (2).

One should note that (RCP) leads to an inconvenience with respect to the second couple of statements: (RCP) would by itself validate (3), so that if (4) is true, so is (3). A Neo-Meinongian is thus led to reinterpret the quantification in (3) and (4) as implicitly restricted to existing objects. Otherwise we could not think of a situation in which (4) is true but (3) is not.[12]

The real problem, however, is that Neo-Meinongianism is not in a position to make sense of the inference from (1) to (2). As far as I can see, a Neo-Meinongian philosopher has only one option to rescue this inference. He has to sacrifice (1) and reinterpret it as really meaning 'everything which is red does not exist'. In other words, a Neo-Meinongian philosopher has to take the superficial structure of (1) to be misleading, since what is really expressed should be formalized as follows (thus abandoning 1** for 1***) (read Ux as the Neo-Meinongian universal quantification):

(1***) $Ux(Rx \supset \sim E!\, x)$

Indeed, it is clear that this reading of (1) would vindicate both inferences, from (1) to (2) and from (2) to (1). This, however, dramatically relativizes the advantage of Neo-Meinongianism vis-à-vis Fregeanism: the former, exactly as the latter, is forced to reinterpret (1) and extract an allegedly deeper logical form to make sense of the raw intuition. The crucial selling point, stressed both by Meinong and Neo-Meinongians, that their approach does justice to the superfi-

[9] This strategy is very common in the literature. It usually takes the following form: some x does not exist implies by (PP) that some existing x does not exist, i.e. a contradiction. A first instance of this strategy is at work in Frege (1883?: 65-6).

[10] See Parsons (1980: 19), Routley (1980: 260-4), and Jacquette (1996: 85-6).

[11] As is well-known, the restrictions are put in place to address the objections by Russell (1905) and Quine (1948).

[12] As Routley (1980: 27) says, "existential loading is a contextual matter".

cial grammatical structure of our language is, at least to some extent, jeopardized.

2.3 The Attempt at a Compromise: A Deflationary Account of Existence

What should we do? Should we give preference to our intuition that predication implies existence and thus amputate (1) as contradictory? Or should we rather deem existence to be an almost perfectly ordinary discriminating property of objects and strongly revise our understanding of (1), so that its real meaning is captured by (1***)? To me, both look like bad solutions: they both are Procrustean in the sense that they force us to amputate some statements (in the case of Fregeanism) or stretch them so as to make them almost unrecognizable (in the case of Neo-Meinongianism). (The reader will remember that the legendary bandit had two opposite ways of torturing his victims, either by amputating their limbs if they did not fit the Procrustean bed, or by stretching them if they did not fill it up.) In other words, neither of the two options is really in a position to do justice to the raw intuition. So, the question should rather be: is there really no better option?

My suggestion will be the following. We can avoid all amputations and reach a streamlined interpretation of the raw intuition by exploring a third possible explanation of existence. To wit, we should abandon the assumption, shared by both accounts, that existence has a nature which we may be searching for, be it that of a pleonastic or of a discriminating property. More generally, following Lewis (1970: 19), we should rather say that there is no connection between the notion of existence and any aspect of the world, be it a property or anything else. Instead, we should consider existence to be a redundant notion, whose meaning is entirely exhausted by the following *existence equivalence schema* (EES) and its negative counterpart, the *non-existence equivalence schema* (NES):

(EES) *n* exist(s) if and only if s_n.

(NES) *n* do(es) not exist if and only if it is not the case that s_n.

As the reader will have noticed, (EES) and (NES) follow the blueprint of the equivalence schemata of the deflationary account of truth and falsity: <*p*> is true if and only if *p* and <*p*> is not true (false) if and only if it is not the case that *p*. Not surprisingly, however, there are crucial differences. First, '*n*' should be understood as a variable for any particular quantified nominal expression, no matter whether in singular or plural form (e.g., 'something red' or 'some red things', respectively). Second, 's_n' should be understood as a variable for the sentence which may be extracted from the nominal expression in question (e.g., 'something is red' or 'some things are red').[13] Finally, a further important difference is that (EES) and (NES) do not involve any metalinguistic shift: there is no device to name linguistic entities, be it sentences or propositions (the square brackets).

One may wonder at this point whether we may apply the equivalence schemata to existential statements with universally quantified nominal expres-

[13] How, then, are we supposed to interpret 'something exists' and 'something does not exist'? The sentence we may extract from 'something' may only be 'something is somehow' or 'something is of some kind'. Thus, (EES) and (NES) yield us, respectively, 'something is of some kind' and 'it is not the case that something is of some kind'. See below, section 5, for further discussion of this pair of statements.

sions such as 'everything red exists' or 'everything red does not exist'. These, however, strike me as ill-formed statements which no one really makes use of. As a descriptivist, moreover, I should point out that I am committed to the thesis that all existential statements with proper names may be cast as quantified statements (roughly put, from 'Pegasus exists' to 'something Pegasizing exists'), so that we do not need any special equivalence schema for such statements.[14]

A further remark is required. The equivalence relation I take to be expressed by (EES) and (NES) is neither an extensional, material one, nor a metaphysical, necessary one. Instead, it must be an analytical equivalence. Only from such a perspective may we say that there is no connection between the notion of existence and any aspect of the world, and that we are, instead, dealing with a redundant notion. Indeed, since nothing expresses a notion of existence on the right-hand side of the equivalence, we may say that the notion of existence is redundant on the left-hand side.

By way of clarification, let us apply (EES) and (NES) to 'something red exists' and 'something red does not exist' (i.e., (1)), respectively:

(5) Something red exists if and only if something (is) red.
(6) Something red does not exist if and only if it is not the case that something (is) red.

In both (5) and (6), 'something is red' is the sentence which may be extracted from the nominal expression 'something red' (I highlight this by putting the sentence-forming device, i.e., the copula, in parenthesis). The deflationist theory I propose is that there is nothing more to be said about existence than what (EES) and (NES) and their instantiations tell us.

Now, from the perspective of our line of reasoning, the crucial advantage of the deflationary account of existence which we have just proposed lies in the streamlined explanation of the raw intuition. Indeed, if we apply (EES) to (2) we get:

(7) It is not the case that something red exists if and only if it is not the case that something is red.

It follows thence that both (1) and (2) are equivalent with another since the application of (NES) to (1) and (EES) to (2) shows that they are both equivalent to a third, identical statement: 'it is not the case that something is red'. We have thus explained their mutual implication.

Another way to state the same point would be to say that (NES) reveals to us why (1) is not really a case of internal negation: (1) is really equivalent to a statement with external negation. One may indeed think of the syntactical predicate 'to exist' as a linguistic device to stress the external negation, in the case of negative statements, and to stress the absence of negation, in the case of affirmative statements. Yet nothing is really added to the content of the statement, since

[14] As far as other natural language quantifiers different from the particular and universal ones are concerned, it seems to me that we may apply (EES) and (NES) to them as well. For instance, the quantified existential statement 'at least one red thing exists' would yield us by application of (EES) 'at least one red thing exists if and only if at least one thing is red'. Or, to take an example suggested by an anonymous reviewer of this paper, the quantified existential statement 'more tigers than lions exist' would yield us by application of (EES) 'more tigers than lions exist if and only if more things are tigers than lions'.

the expression 'to exist' does not refer to a property or nature. This, moreover, is all we need in order to explain the different behavior of (3) and (4), since (3) really confronts us with an internal negation and a predicate which is not merely syntactical but actually adds something to the content of the statement.[15]

The crucial argument developed in this section may now be cast as a trilemma. Let us assume that we want to make sense of the raw intuition. The verb 'to exist' expresses a universal property, a discriminating property, or a redundant concept whose whole meaning is entirely captured by (EES) and (NES). If 'to exist' expresses a universal property, then we have to amputate (1) as contradictory. If 'to exist' expresses a discriminating property, then we have to stretch our language, for (1) can no longer be taken at face value and hides a universal quantification instead. Finally, if the meaning of 'to exist' is entirely captured by (EES) and (NES), then we need neither amputate nor stretch (1). Moreover, if the meaning of 'to exist' is entirely captured by (EES) and (NES), then we need not to revise the general rule according to which internal negation is not interchangeable with external negation, since (1) no longer constitutes a case of genuine internal negation. Furthermore, since I assume that (i) we neither want to amputate nor stretch our language, and (ii) we also have an interest in upholding the general rule that internal and external negation are not interchangeable, we should conclude that the meaning of the syntactical predicate 'to exist' is entirely captured by (EES) and (NES).

Let us now return to any qualms the reader may have with the raw intuition. To such a reader we may say that the deflationary account of existence we have just put forward is not essentially dependent upon endorsing the raw intuition. One may very well not share the intuition that (1) and (2) imply one another. They are, after all, problematic statements, where it is perhaps out of place to rely on intuitions to determine their entailment-relations. Rather, they are statements which should be interpreted in the light of a theory. But then again, even abstracting from the raw intuition, we still have an interest in following the deflationary account of existence. The reason is that going deflationist provides us in any case with a good compromise between Fregeanism and Neo-Meinongianism.

On the one hand, by going deflationist, we avoid the problem of Fregeanism highlighted by Neo-Meinongians: negative existentials of the form 'something such and such does not exist' are no longer contradictory. (I thus assume that we have at least an intuition about the non-contradictory character of such statements.) On the other hand, we equally avoid any Neo-Meinongian distinction between existent and non-existent objects and the epicycles which have to be coupled to this distinction, i.e., the target of the objections raised by Fregean philosophers. It is because the theory concedes something to both contenders that the deflationist approach to existence should be seen as a compromise.

[15] The talk of existence as a merely syntactical predicate that adds nothing to the content of our statements clearly brings to mind both what Kant and Hume had to say about existence. And, indeed, it is tempting to consider both philosophers as defending a kind of deflationism (see Thomasson 2014: 191).

3. Other Raw Intuitions

As remarked at the end of the last section, I am persuaded that the deflationary account of existence is independent of a pre-theoretical endorsement of the raw intuition. But what I cannot and do not want to say is that the account is independent of any linguistic intuition: if the account is convincing, it has to be in conformity with other intuitions a given speaker may have. In other words, the application of (EES) and (NES) to existential statements in the vernacular should not lead to counterintuitive results. Or, at least, we must reach a kind of reflective equilibrium between our intuitions and the deflationary account of existence, so that some intuitions support the theory, while the theory itself should help establish other intuitions. In section 4, I address some (kinds of) existential statements which may seem, in this respect, especially problematic. In this section, however, I would like to draw attention to other intuitions which seem to support the theory. First, I am going to present the reader with a second raw intuition. Then, I am going to introduce a modal declination of the first raw intuition.

3.1. A Second Raw Intuition

Let us consider the following pair of statements:

(8) Something round and square does not exist.
(9) Something round does not exist.

Here it may very well be the case that the former is true and the latter is false (the implication goes only the other way round). Yet things are rather different with the following pair of non-existential statements:

(10) Something round and heavy is not red.
(11) Something round is not red.

Evidently, the former implies the latter: if (10) is true, so is (11).

How would the Fregean and the Neo-Meinongian approach deal with this further raw intuition? As with the previous one, Fregeanism and Neo-Meinongianism would lead, respectively, to an amputation and stretching of our language. If we follow the Fregean approach, we would simply have to amputate statements such as (8) and (9). If we follow Neo-Meinongianism, on the other hand, we should stretch (8) to 'everything round and square does not exist'. In this case, the inference to (9) would clearly not be allowed. If we follow the third way, instead, we may rescue all our intuitions about these statements and at the same time provide an explanation for the alleged bad behavior of existence. The application of (NES) to (8) and (9) yields us (12) and (13), respectively:

(12) Something round and square does not exist if and only if it is not the case that something (is) round and square.
(13) Something round does not exist if and only if it is not the case that something (is) round.

We may now spell out the reason as to why (8) does not imply (9). As a look at the right-hand side of (12) and (13) will show, the reason is that the falsity of a conjunction does not imply the falsity of the conjuncts. Thus, we have once more seen how the deflationary view of existence fares better than Fre-

geanism and Neo-Meinongianism in, as it were, cashing out a pre-philosophical linguistic intuition.

3.2. The Modal Raw Intuition

The time has come to turn to the more complex case of modal existential statements, i.e., statements that involve both the verb 'to exist' and a modal notion. Some philosophers maintain that these kinds of existential claims are the most challenging (think of Moore's 'this might not exist'). In the present section, I would like to focus on a modal version of the raw intuition. First, let us consider a couple of existential modal statements:

(14) Something that might be red does not exist.
(15) It is not the case that something which might be red exists.

As with the non-modal version of the raw intuition, we may again say that at least some of us share the intuition that there is a strong connection between (14) and (15): they mutually imply one another. Notice, moreover, that this would not be the case if we were dealing with something other than existence, and more precisely what everyone would consider a garden-variety property:

(16) Something that might be red is not round.
(17) It is not the case that something which might be red is round.

Everyone would agree that the truth of (16) would simply have no relevance whatsoever for the truth or falsity of (17), and *vice versa*. I am aware of the fact that this second intuition (the difference in behavior of (14) and (15) *vis-à-vis* (16) and (17)) is perhaps even more problematic than the first one: fewer readers are probably going to share it. Yet, as in the case of the first raw intuition, I kindly ask the reader who does not share this intuition to play along until the end of this section. Now, once more, the question we should ask ourselves is the following: why is it the case that in (14) internal negation is interchangeable with the external one?

The dilemma we are facing takes the following form. One option would again be to get rid of such oddities. This would be the path chosen by a Fregean philosopher. Or, more precisely, this would be the path of actualism, i.e., the modal declination of Fregeanism. As it happens, if we formalize (14) and apply the Fregean definition of existence, we will be stuck with a contradiction (let us assume a possible-worlds semantics with constant domains and no restriction on the accessibility relations between worlds):

(14*) $\exists x(\Diamond Rx \wedge \sim E!x)$

An actualist could only make sense of (15), even though he would consider it partially redundant:

(15*) $\sim\exists x(\Diamond Rx \wedge E!x)$

Again, as in the non-modal setting, a Fregean philosopher has two available strategies: he may either say that (14) is contradictory and should therefore be amputated from our language. Or he may say that (14) is just a misleading formulation of the logical form (15*).

What would the Neo-Meinongian alternative look like? According to Neo-Meinongianism, (14) is no longer a contradiction since it may be formalized as (14**) with a non-existentially loaded quantification and a discriminating property of existence:

(14**) $Px(\Diamond Rx \wedge \sim E!x)$

Furthermore, (15) and its formalization as (15**) would no longer be redundant:

(15**) $\sim Px(\Diamond Rx \wedge E!x)$

But how can Neo-Meinongianism vindicate the fact that (14) and (15) mutually imply one another? If we start again by focusing on the inference from (15) to (14), we may see how (RCP) would again validate the inference. However, we may notice that in a modal setting we do not need such a strong principle. Most crucially, once such a principle is introduced, we would also have to grant the inference from a statement such as (18) to (19):

(18) It is not the case that something that could be a round square exists.

(19) Something that could be a round square does not exist.

This consequence may be unwelcome since it would introduce *impossibilia* in our modal logic. A Neo-Meinongian such as Routley would have no qualms with them (see Routley 1980: 83-95). But others may. Hence, in a modal setting, we may prefer to avoid any (RCP) and endorse possibilism: The inference from (15) to (14) is granted from the plain rationalist assumption that for every consistent set of modal properties we have an object that corresponds to it.

As in the non-modal setting, however, the real problem for Neo-Meinongianism or possibilism is the direction of inference from (14) to (15). True, the same option would of course be available as in the non-modal setting, namely reinterpreting (14) as meaning 'everything that might be red does not exist'. This, of course, would lead to abandoning the straightforward (14**) in favor of (14***):

(14***) $Ux(\Diamond Rx \supset \sim E!x)$

Thus, the Neo-Meinongian (as well as the possibilist) is again forced to sacrifice (14), or more precisely, he has to stretch it to the point that it is no longer recognizable.

Having thus introduced the modal raw intuition and explained the challenge it poses to both Fregeanism and Neo-Meinongianism (or to actualism and possibilism), the stage is set for introducing the deflationary view of existence. The reader will have probably already guessed the thesis that I am going to put forward: the inferences which are at stake in statements about actual existence can easily be accounted for as soon as we apply our equivalence schemata. We thus may move away from (14) and (15) on to, respectively, (20) and (21):

(20) Something which might be red does not exist if and only if it is not the case that something might be red.

(21) It is not the case that something which might be red exists if and only if it is not the case that something might be red.

The mystery as to how internal negation is interchangeable with external negation is now easily dispelled. As in the previous case, the predicate 'to exist' does not express any property or nature. Rather, it may be considered a stylistic device to stress negation in the case of negative statements, or, alternatively, the absence of negation in the case of affirmative statements. This, again, is the crucial difference between existence and roundness: (16) is a genuine instance of internal negation. The advantages of the deflationist approach to existence may thus be confirmed in the modal setting too: as soon as we abandon the premise that the verb 'to exist' really expresses a property or a nature, there is no longer any need to amputate or stretch our language.

But what if the reader does not share the modal raw intuition? As with the non-modal raw intuition, I would argue that he still would have an interest in endorsing the deflationist approach. The reason is that going deflationist provides—once again—a good compromise between the modal cousins of Fregeanism and Neo-Meinongianism: namely, actualism and possibilism. On the one hand, we avoid the problem of actualism that negative existentials of the form 'something which might be such and such does not exist' become contradictory (I thus assume, with Moore, that we have at least an intuition about the non-contradictory character of such statements). On the other hand, we equally avoid the possibilist distinction between existent and non-existent objects and the problems which are coupled to this distinction. Thus, the reader who were to choose this perspective may accept the entailment relations between (14) and (15) in the light of this theory and without having to rely on blind intuitions.

It is worth noticing that the line of reasoning just presented may be declined in tensed contexts as well. Here we would have to say that the statement 'something that was red does not exist' and 'it is not the case that something that was red exists' intuitively imply one another (or, more prudently, one might have an intuition to this effect). Then, the same line of reasoning would lead us to deflationism, regardless of worries we might have about the intuition in question. Deflationism about existence thus opens the path for a compromise between the tensed declinations of Fregeanism and Neo-Meinongianism, i.e., what are sometimes labeled, respectively, as presentism and contingentism.

4. Intentional Statements

Intentional statements give rise to two well-known puzzles: failure of substitutivity of co-referring terms and failure of existential generalization. However, both puzzles are linked to the assumption that there are such things as genuine singular statements in our language. And, since in this paper I am abstracting from singular statements, I will set these two puzzles aside. The challenge raised by intentional statements in the present context is thus a different one and has rather to do with the distinction between *de re* and *de dicto* readings. More precisely, if it can be shown that we have *de re* intentional statements about non-existent objects, this would imply that Neo-Meinongians are right after all: the class of objects may be divided into two classes, namely, existent and non-existent ones.

Let us first consider a couple of intentional statements involving the notion of belief:

(22) Meinong believes that something is a golden mountain.

(23) Meinong believes that something which is a golden mountain does not exist.

One should add that both (22) and (23) are true: historically, Meinong really held those beliefs. Now, Neo-Meinongianism would provide us with a *de re* interpretation of these two intentional statements. Not only (22) and (23) are true, but also (24):

(24) Something is such that Meinong believes that it is a non-existent golden mountain.

A Fregean philosopher, however, would clearly resist such an interpretation. To him, Meinong's belief described in (23) is inconsistent in that it implies by (PP)

that there is something which is a golden mountain and does not exist. Similarly, Meinong's belief in (22) is interpreted as false: it is not true that there is something which is a golden mountain. Thus, even though (22) and (23) are true, this does not mean that (24) is true, as well. Or, in other words, we should reject the *de re* reading.

The deflationist view of existence, finally, is more generous towards Meinong and Neo-Meinongianism because it avoids any reference to (PP). The belief in (23) is not inconsistent. Nevertheless, the belief in (23) contradicts the belief in (22) for the very reason that to say that a golden mountain does not exist is, by (EES), tantamount to saying that it is not the case that something is a golden mountain. Thus, the deflationist view of existence allows for a different, more lenient, diagnosis of Meinong's inconsistency. This diagnosis, however, does enough work to block the *de re* reading of the intentional statements in question: (22) is not about a golden mountain because Meinong's belief that something is a golden mountain is false, and (23) is not about a golden mountain even though Meinong's belief that a golden mountain does not exist is true.[16]

But let us turn to a more challenging example of intentional statements:[17]

(25) Meinong imagines a golden mountain.

This intentional statement I take to be equivalent to (26):

(26) Meinong imagines that something is a golden mountain.

The reason why imagination is more challenging than belief is that in this case it seems that something really is a golden mountain, namely what is imagined by Meinong. Does Meinong not have a golden mountain, as it were, 'in front of his eyes' while imagining it? Moreover, since we all assume that golden mountains do not exist, it seems that we have provided a scenario in which both (27) and (28) are true:

(27) Something is a golden mountain.
(28) A golden mountain does not exist.

Or, in other words, it appears that we are forced to accept a *de re* reading of the intentional statement in question. This may be seen as a decisive argument for Neo-Meinongianism and thus a refutation of the view defended in this paper.

Nevertheless, are we sure we know enough about imagination to draw such a conclusion? For one, the following alternative interpretation deserves to be considered: imagination may be nothing else than the ability of our mind to mimic the perception of something. From such a perspective, one should rephrase (26) as (29):

(29) Meinong's mind mimics the perception of a golden mountain.

This seems to be a rather plausible explanation of imagination, which does not require it to be about non-existent mountains. Instead, what is required is simply

[16] The same applies to Meinong's belief that a golden mountain is golden since this implies the belief that something is a golden mountain. This belief is not about a golden mountain either, because it is false. (Of course, Meinong's belief would turn out to be true if interpreted hypothetically: if something is a golden mountain, then it is golden.)

[17] The example is equivalent to the one by Priest (2008b: 296) of John imagining an ugly monster. Priest considers such examples to be crucial evidence in favor of Neo-Meinongianism.

a sensory experience produced by Meinong's mind that mimics the experience he would have if he saw a golden mountain. Notice, moreover, that one may very well say that something is such that Meinong imagines it to be a golden mountain (thus, we may provide a *de re* reading). However, it is not really the case that it is a golden mountain. In fact, it is just an imitation of it.

The same strategy may be applied to intentional statements of desire. Crane (2013: 131-33) discusses the following example:[18]

(30) I desire an inexpensive bottle of Burgundy.

Since we would all agree that inexpensive bottles of Burgundy do not exist, we would again have an argument that allows us to regard existence as a discriminating property of objects. Yet, again, this conclusion may be too hasty. Indeed, it seems plausible to interpret desires as mental states that need to be grounded in imagination or perception: I desire things that I imagine or perceive and which—while being imagined or perceived—are accompanied by pleasurable feelings. If, then, we apply the same interpretation of imagination that was sketched above, we see how statement (30) does not imply any relation to a non-existent inexpensive bottle of Burgundy. To the contrary, what (30) implies is merely a mental event that mimics the perception of an inexpensive bottle of Burgundy.

Someone may object that this strategy, even if it may be effective in the case of imagination and desires, cannot be applied to other kinds of intentional statements. It clearly cannot be applied to the following example (I am considering a variation on this very common example, which does not suppose—according to the approach endorsed throughout this paper—that Ponce de Leon searched for a definite object):

(31) Ponce de Leon searched for a fountain of youth.

This, and similar examples, however, I take to be rather unproblematic. Every time we search for something, we are simply trying to establish a truth about something: namely, where it is.[19] And, of course, in order to ask ourselves where something is, we have to believe or at least assume this something to be such and such. To return to our example, (31) implies that the famous Spanish explorer believed or assumed that something had the property of being a fountain of youth and that he was simply trying to figure out the truth about another statement: namely, the statement about its exact location. And since both the belief and the assumption that something is a fountain of youth are false, there is no reason to give a *de re* reading of (31). Thus, as far as intentional statements such as (31) are concerned, we should not be misled by the superficial analogy with, for instance, the statement 'John kicks a ball'. Instead, we should pay attention to what we mean by the verb 'to search'.

I would like to stress that in the present section I have not argued for a specific account of imagination, desiring, searching and so forth. What I have tried to point out is simply how certain interpretations are *prima facie* plausible and al-

[18] What follows may also be easily applied to intentional statements about fear, where fear may be understood in an analogous way to desires: we fear things that trigger certain feelings when we imagine or perceive them. Examples of intentional statements involving the notion of fear are discussed by Routley (1980: 35-37).

[19] This approach follows a suggestion by Montague (1969: 175), who regards 'to seek' as abbreviating 'trying to find'.

low us to uphold the deflationist view of existence. At the same time, it is crucial to highlight how the above given analyses of intentional statements do not rely on (PP) and thus (a) do not fall together with a Fregean approach and (b) are not question-begging with respect to Neo-Meinongianism.

5. Deflationism and Meta-Ontology

As addressed at the beginning of the paper, Thomasson (2014) recently defended a version of deflationism about existence, which, moreover, she links to a quietist approach to some ontological debates, such as for instance the existence of numbers. In this section, I would like to draw attention to some crucial differences between the version of deflationism I am advocating and Thomasson's. These differences, however, should not hide the common ground between Thomasson's deflationism and mine: we both share the anti-metaphysical stance according to which it is pointless to search for any deep nature of existence.

If we focus on the theory itself, Thomasson's deflationist approach to existence is characterized by establishing a strong link to deflationism about the semantic notions of truth and reference: "the concepts of truth, reference, truth-of, and existence are all interlinked by trivial rules, and deflationisms about any of these notions stand or fall together" (Thomasson 2014: 198). More precisely, Thomasson sees a strong link between the notion of existence and the notion of reference, which she ties by means of the equivalence schemata '<n> refers if and only if n exists' and '<P> refers if and only if Ps exist' (whereby 'n' stands for any singular term and 'P' for any general term different from existence). Then, via the notion of reference, the notion of existence may be tied to the notion of truth to form what Thomasson labels as a "conceptual circle".

The kind of deflationism defended in this paper is, by contrast, independent of deflationism about truth and reference, which of course may be seen as an advantage (if you are a deflationist about truth and reference) or a drawback (if you are not). In addition, the semantic notions of reference and truth are simply not part of the picture I have presented. This strikes me as a clear advantage of the approach I am defending. In fact, we all have a fairly good understanding of existential claims, but only philosophers are familiar with the semantic notion of truth and, especially, reference.

Turning to the meta-ontological implications, Thomasson (2014: 204-206) explicitly develops her brand of deflationism as providing us with a path to "easy ontology". According to this perspective, some ontological questions may be solved by looking at the world. For instance, to know whether red things exist, we have to rely on our conceptual competence and see whether the concept red refers to anything, i.e., whether we have instances of red things. Furthermore, other, less trivial, ontological questions such as the one targeting the existence of numbers, should be seen as trivial inferences from uncontroversial truths, which do not involve the concept at issue (for instance, from 'there are three cups on the table' to 'the number of cups on the table is three'). Thomasson, thus, broadly follows in Carnap's (1950) footsteps and draws a distinction quite close to the one between internal and external questions to a given conceptual framework.

The deflationism defended in this paper, by contrast, is not motivated by and does not have this kind of meta-ontological implications. True, I would agree that in order to assess the question as to whether red things exist we have

to look at the world (notice, though, that I am not bringing into play the notion of reference). But when, for instance, numbers are taken into consideration, the theory defended here does not prescribe any procedure. We may only say—via the application of (EES)—that something which is a number exists if and only if something is a number. The question whether some things are numbers, however, remains open and a legitimate object of ontological dispute.

What, however, is clearly ruled out by the deflationist view of existence defended here is both a Fregean and a Neo-Meinongian approach to ontology. The answer to the question which defines ontology, namely 'what exists?', should neither be 'everything' nor 'the things which happen to have the property of existing'. These are signs of a misunderstanding of the question. Instead, in order to understand the ontological question correctly, we must look at possible answers to it, as for instance in 'something red exists' or 'numbers exist'. All these answers tell us that something is somehow or of some kind (red, number, etc.). Thus, what the question of ontology really means is: what is of what kind? This should be seen as the result of the application of (EES): we have answered the question as to what exists if and only if we have answered the question as to what is of what kind (in other words, the sentence that has to be extracted from the subject 'what' is 'what is of what kind?'). And, if the outcome of such an investigation is that nothing is of any kind, we may move to the nihilist claim that nothing exists, or, equivalently, that it is not the case that anything exists; if, on the other hand, at least something is of some kind, we may confidently state that at least something exists (in other words, the sentence that has to be extracted from the bare subject 'something' is 'something is of some kind').[20]

6. Conclusions

In the present paper, I have outlined a possible deflationist compromise between Fregeanism and Neo-Meinongianism. According to this approach, the two archenemies are both right in their mutual criticisms: existence is neither a universal nor a discriminating property of objects. The reason is that we should simply abandon the assumption according to which existence is a notion that adds something to the content of a statement. Meinong (1904) is famous for having talked about a prejudice in favor of existence, by which he meant the prejudice according to which the only proper objects of scientific enquiry are existent objects. This paper, on the other hand, has argued against a different kind of prejudice in favor of existence, namely that the verb 'to exist' and its cognates express a substantive notion.[21]

[20] See above, footnote 14. These considerations lead to the following interpretation of (PP): 'if something instantiates a property, then this something exists' yields us, by application of (EES), 'if something instantiates a property, then this something is of some kind'. According to this interpretation, (PP) is certainly true but rather vacuous.

[21] This paper is a strongly revised version of chapters 10, 11 and 13 of my PhD thesis (Bacigalupo 2015). I would like to thank the members of the Jury Arkadiusz Chrudzimski, Claudio Majolino, Francesco Orilia, Juan Redmond and, especially, my supervisor Shahid Rahman for their helpful comments and remarks. I am also very grateful to the audience of the SIFA conference in L'Aquila (3-5 September 2014), where I had the pleasure to present an ancestor of this paper. Finally, I would like to thank the two anonymous reviewers for their careful reading and comments.

References

Bacigalupo, G. 2015, *A Study on Existence* (PhD Thesis), Université de Lille 3.

Bourget, D. and Chalmers, D. 2014, "What Do Philosophers Believe", *Philosophical Studies*, 170/3, 465-500.

Branquinho, J. 2012, "What is Existence", *Disputatio*, 34/4, 575-90.

Carnap, R. 1950, "Empiricism, Semantics, and Ontology", *Revue International de Philosophie*, 4, 20-40; reprinted in *Meaning and Necessity*, 2nd edn., Chicago: University of Chicago Press, 1956, 205-21.

Crane, T. 2013, *The Objects of Thought*, Oxford: Oxford University Press.

Frege, G. 1883?, "Dialog mit Pünjer", unpublished during Frege's lifetime; repr. in Hermes, H., Kambartel, F. and Kaulbach, F. (eds.), *Nachgelassene Schriften*, Hamburg: Felix Meiner, 1969, 60-75; transl. by P. Long & R. White as "Dialogue with Pünjer on Existence", in Hermes, H., Kambartel, F. and Kaulbach, F. (eds.), *Posthumous Writings*, Oxford: Blackwell, 1979, 53-67 (references to the page numbers of the English translation).

Hintikka, J. 1966, "On the Logic of Existence and Necessity I: Existence", *The Monist*, 50, 55-76.

Horgan, T. 2007, "Retreat from Non-being", *Australasian Journal of Philosophy*, 84, 615-27.

Lewis, D. 1970, "General Semantics", *Synthese*, 22, 18-67.

Lewis, D. 1990, "Noneism or Allism?", *Mind*, 99, 23-31.

Lycan, W. 1979, "The Trouble with Possible Worlds", in Loux, M.J. (ed.), *The Possible and the Actual*, Ithaca (NY): Cornell University Press, 274-316.

Jacquette, D. 1996, *Meinongian Logic. The Semantics of Existence and Nonexistence*, Berlin-New York: de Gruyter.

McGinn, C. 2001, *Logical Properties*, Oxford: Oxford University Press.

McLeod, S.K. 2011, "First-Order Logic and Some Existential Sentences", *Disputatio*, 31/4, 255-70.

Meinong, A. 1904, "Über Gegenstandstheorie", in *Untersuchungen zur Gegenstandstheorie und Psychologie,* Leipzig: Barth, 1960, 1-50; English translation in Chisholm, R. (ed.), *Realism and the Background of Phenomenology*, Atascadero: Ridgeview, 1956, 76-117.

Mendelsohn, R.L. 2005, *The Philosophy of Gottlob Frege*, Cambridge: Cambridge University Press.

Montague, R. 1969, "On the Nature of Certain Philosophical Entities", *The Monist*, 53/2, 159-194; reprinted in Thomason, R.H. (ed.), *Formal Philosophy. Selected Papers of Richard Montague,* New Heaven-London: Yale University Press, 1974, 148-87.

Orilia, F. 2010, *Singular Reference: A Descriptivist Perspective*, Dordrecht: Springer.

Parsons, T. 1980, *Nonexistent Objects*, New Haven-London: Yale University Press.

Perszyk, K.J. 1993, *Nonexistent Objects: Meinong and Contemporary Philosophy*, Dordrecht: Springer.

Priest, G. 2005, *Towards Non-Being: The Logic and Metaphysics of Intentionality*, Oxford: Clarendon Press.

Priest, G. 2008a, "The Closing of the Mind: How the Particular Quantifier Became Existentially Loaded Behind Our Backs", *The Review of Symbolic Logic*, 1/1, 42-55.

Priest, G. 2008b, *An Introduction to Non-Classical Logic: From If to Is*, 2nd edition, Cambridge: Cambridge University Press.

Quine, W.V.O. 1948, "On What There Is", *Review of Metaphysics*, 2, 21-38; reprinted in Quine, W.V.O., *From a Logical Point of View*, Cambridge, MA: Harvard University Press, 1953, 1-19.

Routley, R. 1980, *Exploring Meinong's Jungle and Beyond*, Canberra: Research School of the Social Sciences.

Russell, B. 1905, "On Denoting", *Mind*, 14, 479-93.

Russell, B. 1918/1919, "The Philosophy of Logical Atomism", *The Monist*, 28, 495-527; 29, 32-63, 190-222, 345-380; reprinted in Russell, B., *Logic and Knowledge*, London: Allen and Unwin, 1956, 177-281.

Thomasson, A. 2014, "Deflationism in Semantics and Metaphysics", in Burgess, A. and Shermann, B. (eds.), *Metasemantics. New Essays on the Foundations of Meaning*, Oxford: Oxford University Press, 185-213.

Van Inwagen, P. 2008, "McGinn on Existence", *The Philosophical Quarterly*, 58/230, 36-58.

Zalta, E. 1988, *Intensional Logic and the Metaphysics of Intentionality*, Cambridge, MA: MIT Press.

Quine, Naturalised Meaning and Empathy

Maria Baghramian

School of Philosophy
University College Dublin

Abstract

Naturalism is the defining feature of the philosophy of Willard van Orman Quine. But there is little clarity in our understanding of naturalism and the role it plays in Quine's work. The current paper explores one strand of Quine's naturalist project, the strand that primarily deals with a naturalised account of language. I examine the role that Quine assigns to empathy as the starting point of the process of learning and translating a language and argue that empathy, when going beyond the automatic form of mirroring, has an irreducible normative character which does not sit well with Quinean naturalism.

Keywords: Quine, Naturalism, Radical Translation, Empathy

1. Quine's naturalism

Naturalism, a dominant strand in current philosophical thinking in the analytic tradition, is the defining feature of the work of Willard van Oman Quine. However, despite its centrality, or maybe because of it, there is no clarity or consensus in our understanding of naturalism nor of the exact role it plays in Quine's work. The current paper explores one strand of Quine's naturalist project, the strand that primary deals with a naturalised account of language.

We can find several interconnected articulations of Quinean naturalism. Metaphilosophical or methodological Naturalism: Philosophy, according to Quine, should not be seen as an autonomous field of enquiry, rather as science conducted at a higher level of abstraction. Here is one famous quote:

> [...] my position is a naturalistic one; I see philosophy not as an *a priori* propaedeutic or groundwork for science, but as continuous with science. I see philosophy and science as in the same boat—a boat which, to revert to Neurath's figure as I so often do, we can rebuild only at sea while staying afloat in it. There is no external vantage point, no first philosophy (Quine1969b: 126-27).

A second defining feature of Quine's methodological naturalism is that he sees the natural sciences as providing the most reliable methodology for any investigation. This approach "sees natural science as an inquiry into reality, fallible

Argumenta 2, 1 (2016): 25-41
ISSN 2465-2334

DOI 10.14275/2465-2334/20163.BAG

and corrigible but not answerable to any supra-scientific tribunal, and not in need of any justification beyond observation and the hypothetico-deductive method" (Quine 1981b: 72).

Naturalising epistemology: This is probably the most widely discussed strand of Quine's project and while continuous with metaphilosophical naturalism, it further extends its scope by proposing that epistemology should be treated as a branch of psychology. To take one representative passage, Quine says: "Epistemology, or something like it, simply falls into place as a chapter of psychology and hence of natural science. It studies a human phenomenon, *viz.*, a physical human subject" (Quine 1969a: 82). And again, "Naturalism does not repudiate epistemology, but assimilates it to empirical psychology" (Quine 1981a: 75).

Quine's aim, then, is not to abandon epistemology but to assimilate it into the empirical science of psychology and ultimately into neuroscience. Epistemological questions, on this account, do continue to hold their legitimacy but are treated as questions within science, rather than prior to it.

Naturalised Meaning: The most radical and controversial strand in Quine's project is the attempt to naturalise language. Linguistic naturalism is significant because it is founded on Quine's momentous rejection of the analytic-synthetic distinction and it is controversial because it leads to the highly counter-intuitive doctrines of indeterminacy of translation and inscrutability of reference. Its two defining features are the rejection of the very idea of meaning and a commitment to a behaviourist view of language acquisition. Indeed, it is sometimes forgotten that Quine's behaviourism did not so much concern the methodology of psychology but was presented as a pre-requisite of linguistics. In *Pursuit of Truth* he is explicit on this point:

> I hold […] that the behaviorist approach is mandatory. In psychology one may or may not be a behaviorist, but in linguistics one has no choice. Each of us learns his language by observing other people's verbal behavior and having his own faltering verbal behavior observed and reinforced or corrected by others. We depend strictly on overt behavior in observable situations. As long as our command of our language fits all external checkpoints, where our utterance or our reaction to someone's utterance can be appraised in the light of some shared situation, so long all is well. Our mental life between checkpoints is indifferent to our rating as a master of the language. There is nothing in linguistic meaning beyond what is to be gleaned from overt behavior in observable circumstances (Quine 1990a: 37-38).

Meaning too is to be understood in term of observable behaviour and dispositions to behaviour. He tells us:

> […] there are no meanings, nor likenesses nor distinctions of meaning, beyond what are implicit in people's dispositions to overt behavior. For naturalism the question whether two expressions are alike or unlike in meaning has no determinate answer, known or unknown, except in so far as the answer is settled in principle by people's speech dispositions, known or unknown. If by these standards there are indeterminate cases, so much the worse for the terminology of meaning and likeness of meaning (Quine 1969a: 29).

Quine's naturalisation of language, of course, begins with the denial of analyticity. To remind ourselves, if any reminders were needed, Quine's strategy, in the "Two Dogmas of Empiricism" (Quine 1951) was to show that none of the

attempts to characterise the analytic/synthetic distinction, including Kant's, but most notably Carnap's, succeed in providing a noncircular or non-question begging account of this well entrenched distinction. He concludes that the belief "That there is such a distinction to be drawn at all is an unempirical dogma of empiricists, a metaphysical article of faith" (Quine 1951: 37). Quine's rejection of analyticity is of a piece with his scepticism about the very possibility of a theory of meaning, as traditionally understood. In its place, he offers an account of language compatible with behaviourist naturalism. Behaviourism is the preferred route because of its scientific credentials. Meaning, in so far as we can allow it into our scientific discourse, is what a sentence has in common with its translation; and translation is understood in terms of establishing correlations between utterances and non-verbal stimulations (see Quine 1960: 32).

The upshot is that meaning is explicated via manuals of translation constructed by observing the stimulus-responses of speakers engaged in verbal behaviour in specific settings. "We can take the behavior, the use, and let the meanings go" (Quine 1979: 1). The price attached to this pared down empirical approach to language is quite high, for we notoriously have to accept that "Manuals for translating one language into another can be set up in divergent ways, all compatible with the totality of speech dispositions, yet incompatible with one another" (Quine 1960: 27).

Quine's approach is not about translation from other languages only but it also applies to attempts at interpreting a 'home' language. Since, the behavioural model applies equally to children learning their first language, it turns out that language is irredeemably indeterminate and the indeterminacy permeates down to the putatively referential singular terms. Neither meaning nor reference can be pinned down and we are thus left with language as

> a social art which we all acquire on the evidence solely of other people's overt behavior under publicly recognizable circumstances. Meanings, therefore, those very models of mental entities, end up as grist for the behaviorist's mill (Quine 1969a: 26).

2. Norming Quine

Quine's naturalist project, and not just his linguistic naturalism, has been criticised for leaving out the normative elements of meaning, knowing and understanding. The worry is that the descriptive language of science is not sufficient for dealing with domains that rely for their proper functioning on normative concepts as well as normative judgements of what is correct, appropriate or desirable. Jaegwon Kim, for instance, has argued that epistemology cannot be fully naturalised because knowledge itself is a normative concept. If we accept the standard definition of knowledge as justified true belief and the intrinsic normativity of justification, then naturalised epistemology would, in effect, amount to the proposal to eliminate knowledge itself from the theory of knowledge (cf. Kim 1988). Similar points apply to the essential normativity of language. Paul Boghossian, for instance, argues that meaningful expressions have correctness conditions and are in that sense essentially normative. "Suppose the expression 'green' means green. It follows immediately that the expression 'green' applies correctly only to these things (the green ones) and not to those (the non-greens). […] The norma-

tivity of meaning turns out to be, in other words, simply a new name for the familiar fact that […] meaningful expressions possess conditions of correct use" (Boghossian 1989: 513).

Quine, however, does allow for norms within his naturalist approach and tells us that naturalism does not require that we "jettison the normative" completely nor that we should settle simply "for the indiscriminate description of ongoing procedures" (Quine 1986: 664). He goes on to explain the role of norms in his account of knowledge:

> For me normative epistemology is a branch of engineering. It is the technology of truth-seeking, or, in a more cautiously epistemological term, prediction. Like any technology, it makes free use of whatever scientific findings may suit its purpose. It draws upon mathematics […] in scouting the gambler's fallacy. It draws upon experimental psychology in exposing perceptual illusions […]. There is no question here of ultimate value, as in morals; it is a matter of efficacy for an ulterior end, truth or prediction. The normative here, as elsewhere in engineering, becomes descriptive when the terminal parameter is expressed (Quine 1986: 664-65).

Quine also considers norms such as simplicity, fecundity, 'conservatism, or 'the maxim of minimum mutilation'' integral to epistemology but argues that they should be seen as playing an instrumental, rather than a foundational, constitutive or a priori role (cf. Quine 1976: 247, 1995: 49). Such norms are tools for achieving our epistemic goals, in particular the goal of increasing our stock of knowledge; their force however is hypothetical rather than categorical. And although epistemic norms are a part of the "technology of truth seeking" (Quine 1986: 665), both truth and knowledge are still to be understood naturalistically—it is only the mechanisms for their discovery that may be norm-laden not the results of such discoveries.

But what about the normative features of language? Do they play any role in Quine's account of language? The suggestion that language has an essential normative element is not devoid of controversy. Boghossian himself, for instance, has conceded (Boghossian 2005) that it is not easy to spell out the conditions for the correct use of language. It is not clear where exactly are we to locate the normative dimension of language and how we are to cash out the idea of 'correct conditions of use' and yet there is a general consensus that to learn a language involves knowledge of what counts as the right conditions for the use of its various elements.

In looking at the normative dimensions of both learning and interpreting/ translating a language we might reasonably argue that the correct usage of a great deal of language is inseparable from the conditions of the truth and falsity of assertoric sentences; to use an assertoric sentence under 'correct conditions' is to make a true assertion, so the normative force of being correct is inseparable from truth. But even if we grant this line of argument, the connection between normativity and truth is far from obvious. Notwithstanding Timothy Williamson's 'knowledge first' project (Williamson 1996, 2007), the claim that truth is a norm of assertion is cashed out in a variety of ways. One standard way of arguing for the point is to claim that to assert a proposition P is to commit to the truth of p, a related way of expressing the point is to argue that by asserting a proposition P we are aiming, maybe not always successfully, to say something true. But, the claim that truth is the goal or the commitment of assertions does not show that

truth itself, in any intuitive sense, is normative (cf. Baghramian and Hamilton 2010). Quine rightly can argue that, at best, what this line of argument delivers is a hypothetical imperative to the effect that if you do not wish to mislead your interlocutors or, if you want to be genuinely informative, then you ought to aim at truth. Quine would be more than happy to admit to this hypothetical demand for normativity as part of his 'technology of truth-seeking'. As we saw, Quine is willing to allow that norms play a crucial role in epistemology, but only in an instrumental sense. There is, therefore, little reason to think that he would not be willing to assign a similar role to the norm of truth in the linguistic domain. This is evident in Quine's evolutionary account of human cognition, where he famously argues that "Creatures inveterately wrong in their inductions have a pathetic but praiseworthy tendency to die before reproducing their kind" (Quine 1969b: 126) because such creatures do not manage to accumulate and communicate a sufficiently large stock of true beliefs that is essential for their survival. Our survival as a species depends on having beliefs that are, for the most part, true. But the usefulness of truth does not show that sentences have to comply with the norm of truthfulness in order to be meaningful. False sentences are as meaningful as true ones and also have correct conditions of application, for instance under conditions when we intend to mislead our interlocutors.

A second and possibly more promising way of assigning a central role to norms in Quine's account of language is via his theory of radical translation. The radical translator, in Quine's famous thought experiment, attempts to translate a hitherto unknown language into his home language by correlating linguistic utterances of the natives with their behaviour and features of the environment. Quine's line of thought is well known, so a very brief reminder should suffice. Radical translation sets out the conditions for translating the language of a hitherto unknown people without help from dictionaries or bilinguals. The only data that the radical translator has at his disposal are the observable behaviour of the speakers of this unknown language and the forces that he can see impinging on the native's surfaces (cf. Quine 1960: 32-33). According to Quine, the sort of meaning that is basic to translation, and to the learning of one's language, is empirical meaning and nothing more. 'A child learns his first words and sentences by hearing and using them in the presence of appropriate stimulus. These must be external stimuli, for they must act both on the child and on the speaker from whom he is learning. Language is socially inculcated and controlled; the inculcation and control turn strictly on the keying of sentences to shared stimulation (Quine 1969a: 81). Unsurprisingly, this austere naturalist view of language learning does not leave much room for norms as either the presuppositions or the essential features of language. The question facing us now, is whether Quine could provide an adequate view of understanding and communication through language while by-passing the idea that norms are integral to the conditions for successful uses of language.

Quine acknowledges that even within his proposed austere linguistic landscape, the field linguist will incorporate certain normative principles into his manual of translation, the most significant of which is the Principle of Charity. Here is Quine's statement of the Principle:

> The maxim of translation underlying all this is that assertions startlingly false on the face of them are likely to turn on hidden differences of language. This maxim is strong enough in all of us to swerve us even from the homophonic method that

> is so fundamental to the very acquisition and use of one's mother tongue. The common sense behind the maxim is that one's interlocutor's silliness, beyond a certain point, is less likely than bad translation—or, in the domestic case, linguistic divergence (Quine 1960: 54).

And again,

> […] the more absurd or exotic the beliefs imputed to a people, the more suspicious we are entitled to be of the translations; the myth of the prelogical people marks only the extreme. For translation theory, banal messages are the breath of life (Quine 1960: 63).

It would be tempting to argue that the Principle of Charity shows the indispensability of norms, or at least some norms, to all translations. Lance and Hawthorne (1997: 12), for instance, have argued that translation is necessarily normative. Their point will apply equally to Quine's method of radical translation. It is the translator's task to make sense of other speakers and, *pace* Michael Williams (Williams 2006: 99), making sense is a fundamentally normative activity. Hence the temptation to claim that the Principle of Charity, the starting point of the process of translation, carries the normative burden that we are assigning to language learning and understanding, a burden that Quine wished to avoid. But the route to 'norming' Quine is not so simple. Two objections to this line of thought present themselves.

Firstly, the norms of translation assumed by the field linguist, according to Quine, but not to Davidson, are instrumental. They are heuristic devices or prudential constraints, rather than indispensable presuppositions of the very act of interpretation. In this, they parallel the epistemic norms allowed by Quine and discussed above. For Quine, at his most radical best, even the laws of logic are defeasible assumptions and open to revision.

Secondly, and even more importantly, Quine at all times, is seeking to ground the Principle of Charity in empirical considerations. For instance, in "Philosophical Progress in Language Theory", where he also argues strongly for the literal continuity between the natural sciences and philosophy by saying that "Philosophy, or what appeals to me under that head, is an aspect of science" (Quine 1970: 2), he urges that the targets of our translation should be construed as expressing 'plausible messages', and proceeds to give an empirical account of such messages, based on frequency measurements and statistical considerations. So, Quine's naturalistic view of language is not undermined by the normative requirements of the assertoric uses of language nor by the requirements of radical translation. Yet, I think the spectre of normativity, of the sort that would not sit readily with Quine's naturalism, still haunts his arid linguistic landscape. In the remainder of this paper, I try to make a case for this very point.

3. Quine on Empathy

From the very outset of developing the project of linguistic naturalism, it has been obvious to Quine, and not just to his critics, that there was more to learning, interpreting, and translating a language than the simple mapping of basic observation sentences to stimuli. One nagging question facing Quine, as well as his commentators, was how is it possible to know or to establish that speakers and learners are acting on the same stimulus. Davidson in a number of places had tried to

persuade Quine that sameness of meaning can be achieved by accepting the role of distal stimuli, shared by speakers (cf. Davidson 1990). Quine, on the other hand, continued to insist that within a naturalist account of translation, the only class of stimulus suitable for a scientific treatment is the stimulation of nerve endings through the individual speaker's encounters with the world, or what Davidson calls 'proximal stimuli' (*ibid*.). Quine, therefore, continued to locate stimulus meaning at the level of the neural input, rather than the external objects of reference. However, he did acknowledge the force of the criticism that an internal psychological account of individual speakers' patterns of assent and dissent to stimuli does not, by itself, explain how speakers could be assumed to have shared sets of stimuli and thereby a shared language. His solution, in the 1980s, was to postulate an innate shared sense of similarity between speakers as the guarantor of sameness of stimulus in the first instance of observation sentences in the next stages of translation. He argued:

> People have to be in substantial agreement, however unconscious, as to what counts as similar if they are to succeed in learning, one person from another, when next to assent to a given observation sentence. […] Subjects radically at odds in this neural way could never learn observation sentences or anything else from one another. Our training even of a dog, horse, bear, seal, or elephant hinges on a conformity of his inarticulate similarity standards to our own (Quine 1984: 294).

However, very soon Quine had to admit that similarity in patterns of stimulus and response does not guarantee the sameness of stimulus meaning, because, for one thing, there are indefinitely many patterns of similarities and differences between any object and state and we need first to determine the respect in which they are similar or dissimilar. What patterns of similarity and differences we pick up at any occasion would depend on contextual considerations and our interests, so what counts as similar is not exhaustively determined by our shared neuronal makeup but also by the contexts that make such judgements relevant or appropriate. Since judgments of similarity, as Quine admits, are substantially interest-relative, in addition to a shared sense of similarity, an interpreter needs to become attuned to what other speakers consider similar on a given occasion. This is where Quine begins to appeal to a shared capability that would make meaning intersubjectively available. He comes to explain this capability in terms of a shared experience of empathy or the ability, perceptually and epistemically, to put oneself in other person's shoes.

The term 'empathy' first occurs in Quine's *Pursuit of Truth* (1990a, Chs. III and IV) and later in his *From Stimulus to Science* (1995, Ch. VIII). But the basic idea that translation requires the ability to project oneself in the place of another can already be found in *Word and Object* (1962) and even earlier in "The Problem of Meaning in Linguistics" (1953) where he writes:

> But, as the sentences undergoing translation get further and further from mere reports of common observations, the clarity of any possible conflict decreases; the lexicographer comes to depend increasingly on a projection of himself, with his Indo-European *Weltanschauung*, into the sandals of his Kalaba informant. He comes also to turn increasingly to that last refuge of all scientists, the appeal to internal simplicity of his growing system (Quine 1953b: 63).

The attempt to put oneself in someone else's shoe, to try and experience the world from their perspective, is core to Quine's understanding of empathy. In later writings he tends to give the term a somewhat wider scope, but two key underlying ideas that empathy is the ability of a subject to project itself onto the mental states of others or to simulate their mental states are central to his conception. The clearest statement of his thinking comes in 1990a, where he writes:

> Empathy dominates the learning of language, both by child and by field linguist. In the child's case it is the parent's empathy. The parent assesses the appropriateness of the child's observation sentence by noting the child's orientation and how the scene would look from there. In the field linguist's case it is empathy on his own part when he makes his first conjecture about 'Gavagai' on the strength of the native's utterance and orientation, and again when he queries 'Gavagai' for the native's assent in a promising subsequent situation. We all have an uncanny knack for empathizing another's perceptual situation, however ignorant of the physiological or optical mechanism of his perception. The knack is comparable, almost, to our ability to recognize faces while unable to sketch or describe them (Quine 1990a: 42-43).

Quine then explains the role empathy plays in the radical translator's attempts at understanding the native's language. Empathy, he believes, remains the guiding principle of the linguist when he moves beyond perceptions and attempts to project grammatical trends and also understand and interpret more complex sentences as well as mental states. He says:

> Empathy guides the linguist still as he rises above observations sentences through his analytical hypotheses, though there he is trying to project into the native's associations and grammatical trends rather than his perceptions. And much the same must be true of the growing child (Quine 1990a: 43).

He further explains:

> Empathy figures also in the child's acquisition of his first observation sentences. He does not just hear the sentence, see the reported object or event, and then associate the two. He also notes the speaker's orientation, gesture, and facial expression. In his as yet inarticulate way he perceives that the speaker perceives the object or event. When the child puts the sentence to use, there is again a perceiving of perceiving, this time in reverse. The listener, concerned with the child's progress, takes note of his orientation and facial expression. The listener is not satisfied by mere truth of the utterance; the child has to have perceived its truth to win applause (Quine 1995: 89).

Both translation and childhood language acquisition require empathy and in both instances two conditions need to apply: speakers should perceive similar stimuli but also the learner/translator has to perceive that the other speaker is perceiving the same stimuli. An example by Peter Hylton clarifies Quine's point:

> for a child to learn, say, "It's raining" from an adult it is not enough that each of them perceives that it's raining; one of them, at least, must also perceive that the other perceives that it's raining. If the child is to learn this sentence as an observation sentence from the adult then one party or the other—and in practice, presumably, often both—must have the capacity to discriminate not only those occasions

> on which it is raining from those on which it is not but also those occasions on which the other party perceives that it is raining from those on which he or she does not. This holds equally for the linguist engaged in radical translation (Hylton 2007: 336).

The radical translator, as in Quine's standard account, constructs a manual of translation through conjectures built on correlations between the native's utterances, her non-verbal behaviour and the goings on in her immediate environment. He relies on the principle of charity by refraining from ascribing glaring falsehoods and favours "translations that ascribe beliefs to the native that stand to reason or are consonant with the native's observed way of life" (Quine 1990a: 46). He also tends to be weary of "complicating the structure to be ascribed to the native's grammar and semantics, for this again would be bad psychology" (*ibid*.). But, most crucially, he, in the absence of evidence to the contrary, will assume that the 'native's mind is 'much like our own'. In doing so, Quine tells us, the translator will be relying on what he calls "practical psychology" and "the method of his psychology is empathy: he imagines himself in the native's situation as best he can" (*ibid*.).

We can experience empathy with non-human animals as well. Through empathy, we are entitled to conclude that the "cat can believe 'A mouse is in there'. The language is that of the ascriber of the attitude, though he projects it empathetically to the creature in the attitude. The cat is purportedly in a state of mind in which the ascriber would say 'A mouse is in there'" (Quine 1990a: 68). Empathy, then, is the medium through which we ascribe propositional attitudes such as 'perceives that' both to human and non-human animals (cf. Quine 1990a: 69).

To take one more example, Quine invites us to consider the case 'Tom perceives the train is late'. Without any deliberate planning or assembling the evidence available, "One empathizes, projecting oneself into Tom's situation and Tom's behavior pattern, and finds thereby that the sentence 'The train is late' is what comes naturally. Such is the somewhat haphazard basis for saying that Tom perceives that the train is late. The basis becomes more conclusive if the observed behavior on Tom's part includes a statement of his own that the train is late" (Quine 1990a: 63). Peter Hylton (2007) argues that, on Quine's account, the sort of empathy required for learning language is the capacity to perceive what someone else is perceiving because we could not learn language at all unless we had empathic capacities of this perceptual kind. This is certainly right and perceptual empathy is certainly one component of Quine's account, but Quine extends the scope of our reliance on empathy beyond the basic projection of observational states to more complex propositional attitudes such as belief. He explains their similarities and differences:

> When we ascribe a belief in the idiom 'x believes that p', our evidence is similar [to the case of ascribing perception] but usually more tenuous. We reflect on the believer's behavior, verbal and otherwise, and what we know of his past, and conjecture that we in his place would feel prepared to assent, overtly or covertly, to the content clause (Quine 1990a: 66).

Empathy also involves cases where we ascribe complex beliefs rather than just basic observational states to other minded creatures, that is cases where we go beyond the sort of states that we might also ascribe to non-human animals. In

such instances, the radical translator, observing the native's behaviour, puts herself in the native's place and attributes to the native beliefs and other mental states that she would have had if she had been in the native's position, in the native's circumstances and cultural context. According to Quine, this projection of thought and ideas, although more tenuous, shares the same basis as the projection of perceptual states.

Quine treats both the shared sense of similarity and innate feelings of empathy, as 'instinctive' features of human psychology. Evolution has inculcated them in us, he thinks, because without them language and learning from each other would not have been possible. However, Quine's reliance on empathy as a prerequisite of translation has raised serious questions regarding the extent of his continued commitment both to behaviourism and to naturalism. Alexander George, for instance, has argued that the introduction of empathy undermines Quine's behaviourism (George 2000: 21-22). Eva Picardi, on the other hand, has found an ambiguity in Quine's account of empathy, a vacillation between a naturalist Darwinian interpretation of empathy vs. a normative Diltheyan view and claims that Quine helps himself to both (Picardi 2000: 132). The authors, I think, have come close to diagnosing the problem that the introduction of the notion of empathy poses for Quine's naturalism, but I do not think they locate the ambiguity and the resulting tension between Quinean naturalism and the normative elements of empathy correctly. In what follows I propose a somewhat different account of the connection and the possible tension between Quine's linguistic naturalism and the normative features of empathy.

Since Quine's first forays into discussions of empathy, there has been much debate on the topic. Indeed, empathy has become a veritable cottage industry, not just in in philosophy and psychology but also in popular culture. In particular, much attention has been paid in distinguishing between different varieties of the phenomenon. In the context of this paper, of particular interest is the distinction between low level, or basic, and high level, more complex, instances of empathy (e.g. Goldman 2011, Stueber 2006) as well as the distinctions between different varieties of higher level empathy.

Low level empathy is standardly characterised as an innate and automatic ability to mimic or mirror some aspects of the mental and emotional states of other minded creatures. In a seminal paper outlining two routes to empathy, Alvin Goldman proposes a distinction between two cognitive systems, or routes, of empathy, what he calls 'mirroring' and 'reconstructive' routes. 'Mirroring empathy' is a form of interpersonal mental isomorphism. The view is based on findings in neuroscience regarding the so-called mirror neurons. The discovery of mirror neurons (see Iacoboni 2009b: 653-55) has given support to the view that a certain 'mental mimicry' or mirroring is experienced by humans, as well as by some other animals, usually at a subconscious level and experience that is essential for both learning from each other and for establishing social communicative ties. Goldman in fact believes that the discovery of mirror neurons provides incontrovertible evidence that low level, basic empathy could be defined as isomorphism or matching conditions between the mental states or experiences of individuals. He writes:

> Since the discovery of mirror neurons and mirroring processes, however, there is much less room for skepticism. There is little doubt about the existence of the processes through which patterns of neural activation in one individual lead, via their

> observed manifestations [...], to matching patterns of activations in another individual (Goldman 2011: 33).

The emotion of disgust is one well-studied case. Evidence from fMRI studies shows that observing a face expressing disgust produces mental mimicry, or empathy in the observer (Wicker *et al*. 2003). When Quine claims, as we saw above, that the "perception of another's unspoken thought" by means of instinctive empathy is "older than language" or that "an infant just a few days old responds to an adult's facial expression, even to imitating it by the unlearned flexing of appropriate muscles" (Quine 1995: 89), his views seem to be in line with, if not a precursor to, the mirroring route to empathy.

Empathy, understood in term of an innate ability of mimicry, is not normative in any interesting sense, for it operates at a pre-conscious level and is a non-linguistic or pre-linguistic stratum of cognition. Both Alexander George and Eva Picardi fail to take note of this point and do not acknowledge the non-normative character of low level empathy. If Quine were to rely on low level empathy only as a precondition for establishing sameness of observations sentences, then the charge that he is introducing a normative element to his pre-requisite of translation will not stand and his linguistic naturalism will remain unscathed.

The second route to empathy is not purely instinctive or automatic. Goldman calls this higher form of empathy 'reconstructive empathy', but the labels 'perspective taking' and 're-enactive empathy' have also been used (Stueber 2006).[1] Contrary to automatic mirroring, higher empathy is a conscious, reflective process, akin to feeling attuned with the mental states of others. One of its core functions is to ascribe mental states to others, something that goes beyond the more basic sharing of similar perceptual contents. This function itself can take different routes and, as we will see, is performed in at least three different ways. It is this type of empathy, I contend, that goes beyond the natural and inevitably invokes norms.

Quine's thinking about empathy seems to encompass both varieties. When Quine characterizes empathy as the "perception of another's unspoken thought" by means of instinctive empathy and claims that empathy is "older than language" (Quine 1995: 89) or when he talks of "an uncanny knack for empathizing with another's perceptual situation" (Quine 1990a: 42), then his focus is on basic or low level empathy. And when Quine writes

> Empathy is instinctive. Child psychologists tell us that an infant just a few days old responds to an adult's facial expression, even to imitating it by the unlearned flexing of appropriate muscles. Dogs and bears are believed to detect fear and anger in people and other animals, perhaps by smell (Quine 1995: 89)

he seems to be thinking of the basic ability of mirroring that has been attributed to specialised mirror neurons. To the best of my knowledge, Quine does not refer specifically to the then very recent discovery of mirror neurons, but I think it is safe to suggest that he would have indeed welcomed this development and would

[1] The debates about different forms of empathy have been conducted largely independently of Quine who clearly was a pioneer in the field. Stueber (2006: 212), however, does cite Quine approvingly.

have seen it a vindication of some of his claims.[2] However, Quine does not seem to think that low level empathy is sufficient for learning or translating a language for in various places he seems to be defending different versions of higher level empathy. To reiterate, higher level empathy is needed when the radical translator goes beyond the perceptual level and attempts to attribute beliefs and desires to the subject. An intimation of this view is evident in passages where Quine calls 'empathy' the method of practical psychology (cf. Quine 1990a: 46)—or what in the literature is generally known as 'folk psychology'—and equates it with the ascription of propositional attitudes to minded creatures. Here Quine's focus is on what in contemporary literature is called 'mind-reading', the ascription of propositional attitudes to others. He seems to be thinking about complex and higher level empathy when, in the passage quoted above (Quine 1990a: 43), he argues that empathy rises above observation sentences to cover native's associations. In these instances, the field linguist "observes the native, hears what the native says, and sees the situation. *He empathizes, puts himself in the native's place*" (Quine and Tomida 1992, emphasis added). In this and other similar passages, Quine seems to be thinking of empathy as something similar to Goldman's perspective taking rather than automatic mirroring. He also warns that when we project ourselves into the minds of others, the "farther we venture from simple discourse about familiar concrete things […], the farther apart the checkpoints tend to be spaced" (Quine 1987a: 28), a position that echoes Goldman's view that higher level empathy is more "effortful and constructive", but less reliable than the low level automatic empathy (Goldman 2011: 30).[3]

Eva Picardi locates the ambiguity in Quine's account of empathy and the pull towards a normative account of interpretation in Quine's failure to distinguish between cases where the empathetic translator tries to figure out what the translator himself would do if he were in the native's place and those cases where he tries to find out what he would do if he were the native. The first reading, Picardi argues, is normative while the second, by appealing to imagination, moves away from behaviourism as classically understood (Picardi 2000: 132), so, on both readings, Picardi argues, Quine forfeits a purely naturalist, behaviourist based account of interpretation and language-learning.

Picardi, I believe, is right in pointing to the absence of finer grained distinctions in Quine's discussions of empathy, but I think, when it comes to the role of what I have called 'higher level empathy', Quine's views could be disambiguated more successfully through Bateson's (2011) distinction between three types of higher level empathy: 1. Intuiting or projecting oneself into another's situation (i.e., simulation); 2. Imagining how another is thinking and feeling not only based on what the other says and does but also based on our own knowledge of other's character, values and desires (what Bateson calls 'imagine other'); and 3. Imagining how one would think and feel in the other's place, 'imagine-self' (or perspective taking). More crucially, contra Picardi, I would like to argue that all three versions of higher empathy involve the exercise of imagination and also rely on

[2] The first paper discussing the functioning of what came to be called 'mirror neurons' was published in *Experimental Brain Research* (91, 1, 1992: 176-80). An earlier article, 1988, in the same journal (71: 491-507) discussed experiment on macaques—area F5. The term 'mirror 'neuron' was first used in an article in *Brain* (119, 2, 1996: 593-609).

[3] Goldman considers Quine's approach in Quine 1990 as an armchair version of the empirically grounded simulationist approach that he has defended (Goldman 2013: 171).

normative presuppositions. In empathy 1, the exercise of empathy involves imagining the circumstances and situations that other people may face. The empathiser "imagines himself in the native's situation as best he can" (Quine 1990a: 46) and tries to decide what she would do, feel or believe if faced with such circumstances. In Empathy 2 or 'imagine other', the empathiser tries to imagine and surmise how the other person may feel or think based on what she already knows about the other person (as in Quine 1990b: 158). He imagines what the subject of empathy would do given her character and psychological makeup. In empathy 3, or imagine-self, the empathiser puts herself in the other person's shoe (or "sandals" as Quine would say) or engages in counterfactual thinking regarding what the empathiser herself might do if he was the other person (e.g. Quine and Tomida 1992). All three types of empathy involve a leap of imagination, so that the difference between them can be explained in terms of the content of what is being imagined.

Even more importantly, all three variants of higher level empathy, unlike automatic mirroring, explicitly or implicitly rely on normative judgements. As we have seen through various quotations from Quine, his account of empathy moves from low level, automatic, mirroring or mimicry to complex acts of ascribing cultural and contextually informed beliefs and other propositional attitudes resulting in full-blown psychological interpretations of others. Empathising, at the more complex level, is a normative act, while low level automatic empathy arguably is not. The point becomes more clear if we look at Quine's suggested strategy of radical translation involving empathy. According to Quine, the radical translator has to rely on observations of the 'local folkways' of his subjects of translation. "[He] will try as an amateur psychologist to fit his interpretations of the native's sentences to the native's likely beliefs rather than to the facts of circumambient nature. Usually the outcome will be the same, since people are so much alike; but his observation of the folkways is his faltering guide to the divergences" (Quine 1995: 80). Quine is not very clear on what he means by 'local folkways' but I believe he uses the term in the sense coined by the turn of the 20th century American sociologist William Graham Sumner meaning conventions and "learned behaviour, shared by a social group, that provides a traditional mode of conduct" (*Encyclopædia Britannica*, 2016). In Quine's account, the translator, through empathy 1, engages in psychological conjectures as to what the native is likely to believe in specific circumstances, or alternatively, through empathy 2, imagines what the native would believe or feel, given her psychological states and, finally, in empathy 3, the empathiser project herself or reads herself "into the minds of others" (Quine 1987: 28-29). Norms are involved in these acts of imagination and counter-factual thinking because empathic understanding is achieved not just within the context of a physical environment but also within the culturally informed web of beliefs, which have a cultural context and would follow norm infused conventions. Folkways, by definition, are imbued with norms; to imagine what one would believe, do or feel in the background of a folkway or culture will inevitably involve judgements about what the native *ought* to believe based on judgements of reasonableness either by the field linguists' lights or by the standards of what the appropriate beliefs are in the context of the natives' way of life. Whichever of higher level empathic routes 1-3 the radical translator adopts, she imbues her translation of the native's utterances with normative judgements, for she is making decisions about what, all things being equal, the appropriate beliefs and other mental states for the native are, i.e. what it is that the subject should be thinking, believing, entertaining in specific circumstances.

A similar point applies to the calculation of what is that the native is 'likely to believe or do'. In situations of radical translation, Quine claims,

> The translator will depend early and late on psychological conjectures as to what the native is likely to believe. This policy already governed his translations of observation sentences. It will continue to operate beyond the observational level, deterring him from translating a native assertion into too glaring a falsehood. He will favor translations that ascribe belief to the native that stand to reason or are consonant with the native's observed way of life (Quine 1990a: 46).

Such informal probability assignments and conjectures often involve assumptions about the rationality of the other person as well further conjectures about their beliefs regarding what ought to be done in specific circumstances. What the translator maximises, according to Quine, is not truth or agreement with his subject, as Davidson had claimed, "but psychological plausibility according to our intuitive folk-psychology", and he insists that "the folk-psychology involved is very much a matter of empathy" (Quine 1990b: 158).[4] But plausibility is a norm governed idea involving assumptions about what is right or appropriate to believe under specific circumstances. As Putnam might have said, cut the empathic pie any way you like, when it comes to higher level empathy, it is difficult to see how a purely naturalist account would suffice.

If the above is correct, then Quine in his later work introduces a norm-governed, and in that sense a non-natural, component to radical translation. Quine might object that even the so-called 'high level empathy', like its low level counter-part, will be shown to have neurological underpinnings and should therefore be understood in naturalistic terms. I have no doubt that this conjecture is correct; nothing performed by the human mind is free of neurological underpinnings. Indeed, according to Goldman the higher level empathy appears to involve a network of neuronal connections "dedicated to shifting perspective from the immediate environment to an alternative situation" (Goldman 2011: 39). But conceding this point does not affect the normative elements of higher level empathy, just as finding neurological underpinnings for our dispositions to behave morally would not render ethics non-normative.

Could Quine rely solely on the low level automatic mirroring account of empathy in explaining language learning and translation? I think the answer, from Quine's own perspective, has to be in the negative. Mimicking or mirroring, at best, gives the language learner the entry point for acquiring the rudiments of language at observational level, the simplest cases of stimulus and response. But Quine admits that convergence on observation would not enable the language learner to understand and translate complex linguistic communications. There is more to language than simply repeating what other speakers say or reacting in similar ways to similar stimuli. To use and understand a language is to be able to apply it, in appropriate ways, to completely novel circumstances. Mimicking or mirroring the language use of our interlocutors will not deliver the creativity and productivity that the use of language requires.[5]

Could Quine argue that the normative features of empathy are yet another version of instrumental norms that he allows in his naturalized account of

[4] This passage in Quine came to my attention from Zanet 2012: 407.
[5] See for instance Chomsky 1966.

knowledge? This particular escape route is not easy to negotiate either. The normative judgments involved in empathic interpretation, the assignment of propositional attitudes to others, do not have the requisite hypothetical form of instrumental reasoning, nor can they be seen as mere tools for achieving specified epistemic goals; rather they are, as Quine outlines them, part of the conditions for the very act of radical translation and interpretation. They are the starting points of the very endeavour to understand and learn a language, an endeavour that marks us off from other animal species. Quine's naturalist account of language then does not escape the need for normative grounding.[6]

References

Baghramian, M. and Hamilton, E. 2010, "Relativism and the Norm of Truth", *Trópos*, III, 1, 31-47.

Barrett, R.B. and Gibson, R.F. (eds.) 1990, *Perspectives on Quine*, Cambridge, MA: Blackwell.

Bateson, D. 2011, *Altruism in Humans*, Oxford: Oxford University Press.

Boghossian, P. 1989, "The Rule-Following Considerations", *Mind*, 98, 507-49.

Boghossian, P. 2005, "Is Meaning Normative?", in Nimtz, C. and Beckermann, A. (eds.), *Philosophy—Science—Scientific Philosophy*, Padernborn: Mentis, 205-18.

Chomsky, N. 1966, *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*, New York: Harper & Row.

Davidson, D. 1990, "Meaning, Truth, and Evidence", in Barrett and Gibson 1990, 68-79.

Goldman, A. 2011, "Two Routes to Empathy: Insights from Cognitive Neuroscience", in Coplan, A. and Goldie, P. (eds.), *Empathy: Philosophical and Psychological Perspectives*, Oxford: Oxford University Press, 31-44.

Goldman, A. 2013, *Joint Ventures: Mindreading, Mirroring, and Embodied Cognition,* Oxford: Oxford University Press.

Guttenplan, S. (ed.) 1975, *Mind and Language*, Oxford: Oxford University Press.

Hahn, L.E. and Schilpp, P.A. (eds.) 1986, *The Philosophy of W.V. Quine*, La Salle, IL: Open Court.

Iacoboni, M. 2009a, *Mirroring People: The science of Empathy and How We Connect to Others*, New York: Picador.

Iacoboni, M. 2009b, "Imitation, Empathy, and Mirror Neurons", *Annual Review of Psychology*, 60, 653-70.

George, A. 2000, "Quine and Observation", in Orenstein, A. and Kotatko, P. (eds.), *Knowledge, Language and Logic: Questions for Quine*, Dordrecht: Kluwer, 21-47.

Hylton, P. 2007, *Quine*, New York: Routledge.

[6] An earlier and much shorter version of this paper was delivered at the The First European Pragmatism Conference: The Relevance of American Philosophy, September 19-21, 2012, conference in Rome. I would like to thank the organisers of that conference and my co-panellists, James O'Shea, Michael Williams and in particular Kenneth Westphal who was generous with his comments. Over the years I have also benefited greatly from discussions and correspondence on the topic of empathy with Meline Papazian. This paper owes a great deal to her ongoing research on empathy.

Kim, J. 1988, "What is 'Naturalized Epistemology'?", in Tomberlin, J.E. (ed.), *Philosophical Perspectives*, 2, Atascadero, CA: Ridgeview, 381-406.

Lance, M.N. and O'Leary-Hawthorne, J. 1997, *The Grammar of Meaning: Normativity and Semantic Discourse*, Cambridge: Cambridge University Press.

Picardi, E. 2000, "Empathy and Charity", in Decock, L. and Horsten, L. (eds.), *Quine: Naturalized Epistemology, Perceptual Knowledge and Ontology*, Amsterdam: Rodopi, 121-34.

Quine, W.V.O. 1951, "Two Dogmas of Empiricism", *Philosophical Review*, 60: 20-43.

Quine, W.V.O. 1953a, *From a Logical Point of View*, Cambridge, MA: Harvard University Press.

Quine, W.V.O. 1953b, "The Problem of Meaning in Linguistics", in Quine 1953a, 47-64.

Quine, W.V.O. 1957, "The Scope and Language of Science", *British Journal for the Philosophy of Science*, 8, 1-17.

Quine, W.V.O. 1960, *Word and Object*, Cambridge, MA: MIT Press.

Quine, W.V.O. 1969a, *Ontological Relativity and Other Essays*, New York: Columbia University Press.

Quine, W.V.O. 1969b, "Natural Kinds", in Quine 1969a, 114-38.

Quine, W.V.O. 1970, "Philosophical Progress in Language Theory", *Metaphilosophy*, 1, 2-19.

Quine, W.V.O. 1974, *The Roots of Reference*, La Salle, IL: Open Court.

Quine, W.V.O. 1975a, "On Empirically Equivalent Systems of the World", *Erkenntnis*, 9, 313-28.

Quine, W.V.O. 1975b, "The Nature of Natural Knowledge", in Guttenplan 1975, 67-81.

Quine, W.V.O. 1975c, "Mind and Verbal Dispositions", in Guttenplan 1975, 83-95.

Quine, W. V.O. 1976, *The Ways of Paradox and Other Essays*, Cambridge, MA: Harvard University Press.

Quine, W.V.O. 1979, "Use and Its Place in Meaning", *Studies in Linguistics and Philosophy*, 3, 1-8.

Quine, W.V.O. 1981a, *Theories and Things*, Cambridge, MA: Harvard University Press.

Quine, W.V.O. 1981b, "Five Milestones of Empiricism", in Quine 1981a, 67-72.

Quine, W.V.O. 1984, "Relativism and Absolutism", *The Monist*, 67, 293-96.

Quine, W.V.O. 1986, "Reply to Morton White", in Hahn and Schilpp 1986, 663-65.

Quine, W.V.O. 1987a, *Quiddities: An Intermittently Philosophical Dictionary*, Cambridge MA: Harvard University Press.

Quine, W.V.O. 1987b, "Indeterminacy of Translation Again", *The Journal of Philosophy*, 84, 1, 5-10.

Quine, W.V.O. 1990a, *Pursuit of Truth*, Cambridge, MA: Harvard University Press.

Quine, W.V.O. 1990b, "Comment on Harman", in Barrett and Gibson 1990, 158.

Quine, W.V.O. 1991, "Two Dogmas in Retrospect", *Canadian Journal of Philosophy*, 21, 265-74.

Quine, W.V.O. 1995, *From Stimulus to Science*, Cambridge, MA: Harvard University Press.

Quine, W.V.O. 1996, "Progress on Two Fronts", *Journal of Philosophy*, 93, 159-63.

Quine, W.V.O. and Tomida, Y. 1992, Interview of Quine (https://sites.google.com/site/diogenesphil/quine-tomida).

Stueber, K.R. 2006, *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences*, Cambridge, MA: MIT Press.

Wicker, B. *et al*. 2003, "Both of Us Disgusted in *My* Insula", *Neuron*, 40, 3, 655-64.

Williams, M. 2006, "Realism: What is Left?", in *Truth and Realism*, Greenough, P. and Lynch, M.P. (eds.), Oxford: Oxford University Press, 77-99.

Williamson, T. 1996, "Knowing and asserting", *Philosophical Review*, 105, 489-523.

Williamson, T. 2007, *The Philosophy of Philosophy*, Malden, MA: Blackwell.

Zanet, G. 2012, "Quine and the Contemporary Debate on Mindreading", *Disputatio*, IV, 32, 395-412.

True but Also Not True

Stefano Boscolo and Giulia Pravato***

**University of Palermo*

***University of Bologna*

Abstract

We present three ways of expressing a possible interpretative uncertainty of the truth predicate: ambiguity, context-sensitivity and semantic indeterminacy. Next, we examine Kölbel (2008)'s pluralist view that "true" is ambiguous between a substantialist concept and a deflationist concept, and that it is ambiguous as the word "dog" is between "male dog" and "canine". Our main goal is to show that Kölbel's thesis does not withstand empirical scrutiny in the sense that "true" fails most of the well-established tests for ambiguity (conjunction-reduction, contradiction, and ellipsis). In addition, we reformulate Kölbel's thesis by saying that "true" may be context-sensitive between a substantialist concept and a deflationist concept, and then we run Cappelen and Lepore (2004)'s inter-contextual disquotation test in order to show that "true" does not display that sort of context-sensitivity. In conclusion, we offer a diagnosis of Kölbel's thesis failure, and advance some possible developments.

Keywords: truth, pluralism, ambiguity, context-sensitivity

1. Introduction

Alethic pluralism is the view that truth requires different treatments for different domains of discourse. Accordingly, the subject matter we are talking about determines what notion of truth is in place. The intuition behind alethic pluralism is that we do not seem to appeal to the same notion of truth across different domains of discourses such as mathematics and morality. This may also account for the widespread disagreement among philosophers on what the nature of truth ultimately is (correspondence, coherence, deflationary, etc.). But given the distinction between concepts and properties, it makes sense to wonder whether there is a single concept of truth but different truth properties, or there are different truth concepts, each of them matching different properties. Wright (1992) argues that there is a single concept of truth that is specified by a list of platitudes about truth, and that different truth properties satisfy that concept in different regions of discourse. On the other hand, Kölbel (2008) argues that truth is split into different concepts which, in turn, are associated with different properties. Unlike Wright's pluralism, which is widely discussed in the literature, Kölbel's has not received

Argumenta 2, 1 (2016): 43-54
ISSN 2465-2334

DOI 10.14275/2465-2334/20163.BOS

much attention yet. Our goal is to examine whether or not Kölbel's pluralism is tenable by looking at how ordinary speakers use the truth predicate.

Kölbel claims that the predicate "true" displays an interpretative uncertainty in natural language between two concepts: a deflationary concept and a substantialist concept. The former, TRUTH-D, is exhausted by the equivalence schema "<p> is true iff p"; the latter, TRUTH-C, involves a relation between truth-bearers and objective states of affairs. At first glance, it is unclear what sort of interpretative uncertainty Kölbel has in mind, as there are at least three ways of fleshing out his claim: the predicate "true" could be ambiguous, context-sensitive or indeterminate. Although these attributes are closely related to one another, as they all point to a lack of certainty, it is nonetheless possible to set them apart. Our inquiry mainly addresses the thesis that "true" is ambiguous and touches on the context-sensitivity alternative.[1]

Let us first introduce some basic terminology. Consider the pun "burying a treasure by the river, Barbarossa is putting his money in the bank". The pun is admittedly funny when one recognizes that "bank" is ambiguous between an organization that provides financial services and the side of a river. Ambiguity is generally defined as a matter of two or more lexical entries that correspond to the same word (e.g. financial-bank and river-bank). Sometimes it is also useful to draw a further distinction between two forms of ambiguity: homonymy and polysemy. A word is homonymous if it has one single phonological form and two separate dictionary entries. For example, the word "coach" has two unrelated meanings: one is trainer; the other is bus. On the other hand, a word is polysemic if it has one single phonological form and two distinct but related meanings. A polysemic word is "face", which can either refer to the part between the forehead and the chin or to the forward part of a clock. Yet the two meanings are conceptually related in that they refer to the front of an object.

It may be hard to tell whether a term is homonymous or polysemic. The word "bank" is clearly homonymous between river and financial institution; nonetheless, it may be considered polysemic between financial institution and relying on someone (as in the expression "you can bank on me") because of the underlying theme of security. For the sake of simplicity, we will lump homonymy and polysemy together and consistently use the term "ambiguity".

Context sensitivity is variability in content due to changes in the context of utterance without any changes in word usage. For example, the personal pronoun "I" is context-sensitive because it shifts reference depending on who is uttering it; but notice that "I" is not ambiguous. Looking up "bank" in the dictionary, we notice two distinct entries that correspond to it. And we conclude from that evidence that the word "bank" must be ambiguous. The personal pronoun "I", by contrast, has only one single lexical entry. To put it another way, "I" has one single lexical entry regardless of whoever is uttering it. Ambiguity, roughly speaking, is a property of the meaning of terms on their own, whereas context-sensitivity is determined by a mix of linguistic facts on the one hand, and non-linguistic facts about possible contexts of utterance on the other.

[1] We mainly examine the ambiguity thesis because Kölbel himself (2008) considers the truth predicate as ambiguous. To be fair, Kölbel also suggests that "true" might display pragmatically ambiguity or context-sensitivity, although he seems to be in favor of syntactic ambiguity (2008: 369). We would also like to emphasize that Kölbel does not take into account "true" as indeterminate.

Keeping the distinction between ambiguity and context-sensitivity is a thorny issue when the dictionary cannot be a reliable tool. Philosophical disputes about whether a term is ambiguous or context-sensitive cannot be settled by merely appealing to dictionaries. Just think about the long-lasting disagreement over words such as "truth", "exist" or "real". Hence we need a method of telling whether a purported term is ambiguous, context sensitive or neither. As will be shown, sections § 3 and § 4 are aimed at presenting such a method and applying it to our ordinary usage of the predicate "true".

Both ambiguity and context-sensitivity should be distinguished from indeterminacy. We say that a term is (semantically) indeterminate if our inability to assess an instance of it would persist even if we had all the relevant information. Consider the sentence "Smith is bald" where Smith is a borderline case. "Bald" is indeterminate because our inability to assess whether or not Smith is bald would persist even if we knew the exact number of Smith's hair. To give another example, we know that *there is* a wealthiest poor person, but we do not know *who* the wealthiest poor is. Basically we know an existential sentence to be true without knowing any instance of it to be true.

If terms like "bald" or "wealthiest poor person" are semantically indeterminate, having complete knowledge of the history of the world (past, present and future) is not sufficient to address questions such as "is Smith bald?" or "who is the wealthiest poor person?" Compare the semantic indeterminacy of "bald" with the ambiguity of "bank". "Bald" does not have clear-cut extension, whereas "bank" *does* have it. This is because we can easily recognize whether something is either a river or a financial institution given all the relevant information. To disambiguate a word we need to add additional information to the context of utterance, whereas there is no fact of the matter to be known in the case of indeterminacy. Notice that context-sensitivity is also quite different from indeterminacy. "Bald" involves blurred conditions of applications, so that our thoughts and practice do not determine the truth-conditions of borderline cases where "bald" occurs. In the case of context-sensitivity, on the contrary, we can determine the truth-conditions of sentences such as "I'm eating an ice-cream right now" when the speaker is clear from the context of utterance.

It may seem that semantic indeterminacy and vagueness describe the same phenomenon, so that every occurrence of "semantic indeterminacy" simply stands for "vagueness". But this is not the case. In fact, three features are typically associated with vagueness: the presence of borderline cases, the lack of sharp boundaries and the sorites-susceptibility. Unlike vagueness, semantic indeterminacy is only characterized by the presence of borderline cases without boundarylessness. Consider Fine's (1975: 266) example of a stipulated definition of "nice1": (a) n is nice1 if n> 15; (b) n is not nice1 if n< 13. Because it is impossible to determine whether or not 14 falls under that predicate, "nice1" is semantically indeterminate if n=14. However, "nice1" is not vague. Should it be vague, it would lack sharp boundaries. Nor is "nice1" affected by sorite paradoxes.

We have shown that ambiguity, context-sensitivity and indeterminacy are distinct notions. Saying that "true" displays an interpretative uncertainty thus requires further elucidation. We are now addressing the main question of the paper: in what sense might "true" be ambiguous?

2. Kölbel's Ambiguity Thesis

The sort of ambiguity we are going to examine is endorsed by Kölbel (2008). Kölbel argues that our ordinary usage of "true" expresses a deflationary concept (TRUTH-D) on some occasions, and expresses a substantialist concept (TRUTH-C) on other occasions. The deflationary concept is exhausted by some variant of the equivalence schema (e.g. the proposition that p is true iff p), whereas the substantialist concept is a "metaphysically interesting concept worthy of further analysis" (Kölbel 2008: 359). More specifically, TRUTH-C is defined by the principle that truth is objective, where objectivity is cashed out in terms of faultless disagreement. A truth-bearer is objective if and only if "disagreement about it cannot, as an a priori matter, be faultless" (Kölbel 2008: 376).

Let us explain Kölbel's view by way of example. Suppose that Sarah and Smith are having a disagreement about whether it is true or false that the voltage induced in a closed circuit is proportional to the rate of change of the magnetic flux it encloses. Since their disagreement is about one of Maxwell's equations, either Sarah or Smith must be wrong. The disagreement in question is not faultless. On the other hand, suppose that Sarah is now quarreling with Smith about whether oysters are tasty or insipid. Because taste judgments are not objective, it is not the case that one of them must be mistaken. Sarah and Smith are thus having a faultless disagreement.

According to Kölbel, competence with the predicate "true" requires knowing that the term expresses a deflationary concept on some occasions of use, and a substantialist concept on some other occasions. In turn, one must accept all the instances of the equivalence schema in order to be competent with the deflationary concept of truth; furthermore, being competent with the substantialist concept requires being acquainted with the notion of objectivity.

Kölbel argues that ordinary speakers are able to disambiguate TRUTH-D from TRUTH-C. As evidence for this claim, he asks us to consider the following two utterances:

(U1) It is true that Chaplin is funny.
(U2) Judgments (propositions, statements, beliefs, etc.) about what is funny cannot be true or false.

Kölbel observes that it is possible for the same speaker to utter both (U1) and (U2) without a change of mind or being confused. For there are two concepts of truth at stake: in (U1) "true" expresses the deflationary concept, which applies to all contents of thought/speech; in (U2) "true" (and "false") expresses a substantialist concept, which only applies to objective contents. In other terms, (U2) says that judgments about what is funny cannot be true-c (or false-c) because they are not objective, whereas "true" in (U1) does not discriminate between objective and subjective contents.

Kölbel's thesis is not just that "true" is ambiguous, but that it is ambiguous in a peculiar way. Unlike "coach" and "bank", which have mutually exclusive meanings, "true" functions as "dog", which has a general understanding ("dog" as canine) and a specific understanding ("dog" as male dog). As "dog" conveys both meanings, so does "true":

For all x, x is a dog-m iff x is dog-c and a male.
For all p, p is true-c iff p is true-d and p is objective.

To sum up, Kölbel's thesis is that "dog" and "true" are likewise ambiguous. What we want to do is to evaluate whether this thesis withstands empirical scrutiny. If that is the case, we should expect "true" to pass the same tests for ambiguity that "dog" also passes. Our general approach consists in putting the predicate "true" in utterances so as to highlight its purported ambiguous features. We shall then consider three well-established tests for detecting ambiguities: conjunction-reduction, contradiction and ellipsis.[2]

3. The Ambiguity Thesis Under Test

Let us start with the test of *conjunction-reduction*. It consists in conjoining two sentences that contain a purportedly ambiguous term, and in showing that the resulting conjunction is zeugmatic. A chain of words is zeugmatic if it must be understood in two different ways in order to make sense. Consider, for instance, the adjective "light" in unambiguous sentences such as (1) and (2):

(1) The colors are light.
(2) The feathers are light.

We build a new sentence by conjoining (1) and (2):

(3) The colors and the feathers are light.

(3) passes the test, for it is clearly zeugmatic. This is evidence that "light" is ambiguous between "not dark" and "not heavy". Consider now the word "exist".

(4) Alghero exists.
(5) Numbers exist.
(6) Alghero and numbers exist.

Regardless of our view on the existence of mathematical objects, (6) is not zeugmatic. This result squares with philosophers' intuition that "exist" is unambiguous. Consider now the following examples where "true" occurs.

(7) That Chaplin is funny is true.
(8) That Chaplin died in 1977 is true.

We have encouraged two readings of "true" as expressing TRUTH-D in (7) and TRUTH-C in (8). But their conjunction does not seem to display any zeugmatic effect:

(9) That Chaplin is funny and that Chaplin died in 1977 are true.

The test seems to drive us to conclude that "true" is not ambiguous in Kölbel's sense. Unfortunately, the matter is a little trickier. Let us put the word "dog" to the test:

(10) Bitches are dogs.
(11) Fido is a dog.
(12) Bitches and Fido are dogs.

Neither (12) strikes us as zeugmatic. So we ought to conclude that "dogs" is unambiguous, which is not the result we expected. What went wrong? The problem is that conjunction-reduction does not seem to work on privative opposites, i.e. when a more general understanding implies a more specific one. Consider the term "lion".

(13) Lionesses are lions.

[2] The orthodox source for these tests is Zwicky and Sadock (1975).

(14) Simba is a lion.
(15) Lionesses and Simba are lions.

Conjunction-reduction fails to detect the ambiguity between feline and male lions. Also, think about the verb "drink", which has a general understanding (drink a liquid) and a specific one (drink alcohol). It is not clear how to emphasize that distinction within conjunction-reduction. Therefore, one may object, the test fails because it cannot display the sort of ambiguity that Kölbel has in mind.

Let us look at another test and see if we get different results. The test of *contradiction* is reliable as evidence to detect ambiguities in privative opposites. Accordingly, an expression is ambiguous if the same string of words can be used to say something that is simultaneously true and false of the same state of affairs. The seeming contradiction should go away as soon as we emphasize the two meanings of the ambiguous term. In this sense, (16), (17) and (18) are all ambiguous:

(16) She was funny [amusing] without being funny [strange].
(17) That bank [river-bank] isn't a bank [financial-bank].
(18) Bitches are dogs [canines] and aren't dogs [male canines].

As the test works on "dog", we can perform it on "true" as well. Consider the sentence

(19) "Chaplin is funny" is true, but it is also not true.

It does not seem that an ordinary English speaker can utter (19) without contradiction. But it is still possible that an ordinary English speaker may not recognize an ambiguity at first sight. After all, even the two understandings of "dog" sound rather unnatural. In other words, what if we specify that in the former instance of true we mean only that Chaplin is funny, whereas in the latter we mean that "Chaplin is funny" is objectively true?

(19') "Chaplin is funny" is true [true-d], but it is also not true [true-c].

Imagine a situation where I am having a conversation with a friend who says, "I watched Modern Times last night. Chaplin is so funny!" I nod in approval and say, "It's true!" Later on, another friend, a professional philosopher this time, comes to me and asks, "I've heard what you said earlier. But do you really believe that Chaplin is funny?" I pause for a second and then reply, "No, I do not believe that taste judgments can be true or false".

The bottom line is that natural language does not display two meanings of "true" unless we distinguish a serious, philosophical context from an ordinary one. That is why, we believe, the contradiction test fails when "true" occurs in *ordinary* speech; this is why we believe that "true" is not ambiguous in Kölbel's sense. But, one may object, what if a philosophical inquiry could reveal what the ordinary speaker is actually committed to? Perhaps the ordinary speaker could be driven to interpret the second occurrence of "true" in (19) as true-c. Assume that we ask an ordinary speaker, "do you think that Chaplin is funny in the same sense that it is true that Alghero is in Sardinia?" It is certainly possible that such an ordinary speaker, after being puzzled, would reply, "it doesn't sound right!" Notice that confusion may arise even after restating the same question on the word "exist". One could insist on the same ordinary speaker asking, "and so does Alghero exist in the same sense that numbers exist?" This question would give the ordinary speaker a hard time as well. At this point, we would need a philosophical

argument to prove that "exist" and "true" are both unambiguous. But this objection, although legit, goes far beyond the aim of the contradiction test. The test neither provides a knock-down philosophical argument nor engages an ordinary speaker in a philosophical debate. It ought to grasp the ambiguity of "true" in utterances without whatsoever philosophical bias. If those utterances sound naïve and not deeply philosophical, then the test is doing its job right. Note that the test does not rule out that there may be *contexts* where the distinction between true-c and true-d holds. But detecting such a distinction is more suitable for a context-sensitivity test (see § 4).

An important caveat: we do not want to suggest that philosophers cannot posit two meanings of the word "true"; but we adhere to the principle that, paraphrasing Grice, *ambiguities are not to be multiplied beyond necessity*. To quote Kripke (1979: 243), "do not posit an ambiguity unless you are really forced to, unless there are really compelling theoretical or intuitive grounds to suppose that an ambiguity really is present".

Let us consider a further test to detect ambiguities. The *ellipsis* test aims to identify impossible conflicting interpretations in sentences of the form *X does/did Y and so does/did Z*. Impossible conflicting interpretations are different readings of the same term that become mutually exclusive once we put them in an elliptical clause. Consider

(20) I went to the bank.

(20) has two conflicting readings. It can mean either that I went to the money institute, or that I went to the river. Now we add an "and so did" clause,

(20') I went to the bank, and so did Bill.

We get two impossible conflicting interpretations. In fact, (20') cannot mean that I went to the money institute (or the river) and that Bill went to the river (or the money institute). These two readings are mutually excluded by (20'). As a result, "bank" must be ambiguous. Notice that "dog" has also two impossible conflicting readings:

(21) I had my dog castrated, and so did Bill.

Suppose that I have a male dog, but Bill has a female dog by the name of Mia. Since castration can be performed only on male dogs—the correct term for females is spaying, or neutering for both males and females—Mia cannot be castrated. As a result, (21) admits impossible conflicting interpretations. "Dog" is therefore ambiguous between dog and male dog.

It is interesting to run the ellipsis test on the word "child". Consider

(22) I adopted a child, and so did Bill.

"Child" can have conflicting interpretations. For instance, it can pick out either a girl or a boy. But these interpretations are not mutually exclusive when we add an "and so did" clause. My child does not need to have the same gender as Bill's one in order for (22) to make sense. It follows that "child" is not ambiguous. At best, it is context-dependent with respect to whether "child" refers to either a boy or a girl.

Now, consider a perverse and horrifying society where children are customarily castrated.[3] The sentence "I castrated a child, and so did Bill" has still conflicting interpretations, boy or girl, but this time they are mutually exclusive. Note

[3] An anonymous referee mentioned this grisly scenario.

that we have not changed the ordinary meaning of "child" but just imagined a scenario where "child" is ambiguous. There are however two important aspects to notice about this imaginary scenario: a) by running a thought experiment, as in the case of a perverse society, we force a change in the predominant context of utterance. Thought experiments have the power to induce ambiguities in a word via an alteration of the context of utterance, and they can be used to prove that a word is context-sensitive or ambiguous under possible scenarios. But we are employing the ellipsis test only to examine the behavior of "child" in ordinary contexts, and so we cannot help ourselves with any thought experiment; b) even if that thought experiment were valid, "child" would not be ambiguous in the same sense as "dog". In fact, "child" would display impossible conflicting interpretations that are polar opposites with respect to a gender feature (i.e. a child who can be castrated and another who cannot),[4] rather than displaying privative opposites (i.e. a general meaning and a specific meaning).

Let us now see how "true" behaves in an ellipsis test. Consider

(23) I think that "Chaplin is funny" is true, and so does Bill.

Suppose that "true" has two conflicting interpretations, namely true-c and true-d. When I say that "Chaplin is funny" is true I simply mean that Chaplin is funny, whereas Bill means that Chaplin is funny and that we cannot have faultless disagreement on Chaplin's funniness. It seems to us that (23) makes sense despite the fact that Bill and I disagree on the objectivity of Chaplin's funniness. Conflicting interpretations of "true" are thus possible, and therefore "true" is not ambiguous. On the assumption that "true" has conflicting interpretations, we can at best conclude that "true" is context-dependent with respect to whether it refers to something that is either objective or subjective. We are going to say more about that assumption in the next section. Of course, it is possible to adopt a generic meaning of "true" and a specific meaning that is entailed by the generic one. But, we stress, this move is a provision rather than the way ordinary speakers commonly use "true" in a statement such as "X is funny" is true.

Here is what we have so far established. We have performed three tests in order to demonstrate that "true" is not ambiguous in Kölbel's sense. Our first conclusion is that the truth predicate fails every test we have presented. This should provide evidence that the truth predicate is not ambiguous in natural language. A corollary of our analysis is that even if "true" were ambiguous in a way that these tests could not detect, it would not have the same type of ambiguity as "dog". In fact, "dog" passes both the contradiction test and the ellipsis test. Again, "true" would not still be ambiguous in Kölbel's sense.

A final caveat: the tests for ambiguities must be handled with care and, in any case, the ambiguity theorist is free to insist that "true" is ambiguous in a manner that is undetectable by the tests. However, in the absence of a better account, Kölbel's proposal stands on shaky grounds.

[4] Two meanings are polar opposites with respect to a semantic feature *F* if they are identical except that the former can be represented as having *F* where the latter without having *F,* or the reverse (Zwicky and Sadock 1975: 6). To give another example, father and mother are polar opposites with respect to a gender feature.

4. A Look at the Context-Sensitivity Thesis

At the end of the last section we allegedly conceded that "true" might have two possible conflicting interpretations. In fact, one may be inclined to weaken Kölbel's thesis by saying that "true" is context-sensitive, in the sense that there are contexts where "true" means true-d and contexts where it means true-c. Before looking at our counter-argument, here is an important methodological proviso. We endorse the principle that if an expression is context-sensitive in English, that is a fact about the English *language*. So the context-sensitivity thesis cannot be disputed on the basis of philosophical arguments.

We want to argue against the context-sensitivity thesis by appealing to Cappelen and Lepore's inter-contextual disquotation test (ICD).[5] The test aims to detect whether or not an expression is context-dependent in ordinary language. An expression *x* is context-dependent iff one can assert that, for some sentence "S" containing *x*, there are false utterances of "S" even if S. For instance, consider the expression "I" in this sentence:

(24) I'm German.

We want to evaluate whether there are false utterances of "I'm German" even if I'm German. The answer is clearly "yes", as there are contexts in which there are false utterances of (24); that is to say, when (24) is uttered by someone who is not German. Consider now the expression "that" in the sentence

(25) That's cute.

Are there false utterances of "that's cute" even if that's cute (said pointing to a kitty cat)? The answer is "yes". Just think about someone who is uttering that expression referring to someone/something that is not cute.

Since "I" and "that" are indexical, it should come as no surprise that such expressions pass ICD. The truth predicate, on the other hand, is more difficult to evaluate. Adopting Cappelen and Lepore's methodology, we ought to build a story (called "context shifting argument" or simply CSA) in which the alleged context-sensitive expression has true utterances while denying an actual use of that sentence.

Consider the word "red". An ICD for red would be as follows: there are false utterances of "apples are red" even though apples are red. To deny that "red" is context-sensitive, there must not be such utterances. Let us look at this CSA (adapted from Cappelen & Lepore 2003: 33):

> Here are some red apples. An apple is red because it has red skin, so those apples have red skin. There are false utterances of "apples are red", not because red apples have changed color, but because the speaker cares about what is inside the apples rather than whether or not they are red.

This argument does not provide a convincing support for the context-sensitivity of red. So we cannot argue for the context-sensitivity of "red" on the basis of that

[5] We are aware that Cappelen and Lepore's work is much controversial insofar as they use their test in order to argue that few purported contextually dependent expressions are such. We are also aware that there are other tests for context-sensitivity; for instance, the agreement-based tests used by Cappelen and Hawthorne (2009). Nonetheless, we would like to note that, unlike ambiguity, there is not a set of standard tests for context-sensitivity. In this respect, we view Cappelen and Lepore's ICD as a worthy attempt.

CSA. Of course, the burden of proof in every ICD depends on the CSA we devise, and on how much the CSA is persuasive. Let us now turn to the predicate "true". Consider

(26) It is true that Chaplin is funny.

We want to evaluate whether there can be false utterances of "it is true that Chaplin is funny" even if it is true that Chaplin is funny. What sort of CSA are we looking for? Kölbel's examples of "true-d" and "true-d" put some constraints on what an appropriate CSA should be. Consider the following CSA:

> Smith is saying that "Chaplin is funny" is true, and by that he simply means that Chaplin is funny. There is a false utterance of "the judgment that Chaplin is funny is true" not because Smith thinks that Chaplin is not funny, but because he also believes that taste judgments are neither true nor false.

Here we are actually telling a story that includes two "target contexts". In an ordinary context, Smith is saying that he believes that it is true that Chaplin is funny, whereas in a philosophical context (when he discusses taste judgments) Smith believes that it is not true that Chaplin is funny. That may be acceptable for Smith, insofar as he is a contextualist about truth, but our intuition on this CSA is that Smith ought to make up his mind. In fact, we are entitled to ask, "OK, but do you believe that Chaplin is funny or not?" What looks bizarre is that Smith is not able to answer a simple question like that without summoning two contexts: one where Chaplin is funny, and one where Chaplin is not funny because taste judgments are neither true nor false. Compare this situation with the case of "I'm German". Even though "I" is context-sensitive, Smith will have no problems in answering the question, "are you German?". Smith will not say, "well, that depends!", and then mention different contexts in which he would employ the word "I". In Cappelen and Lepore's lingo, the problem is that Smith's CSA is an ICSA (impoverished CSA), an argument where the alleged context-sensitive expression is neither asserted nor denied to describe a target context. In our example, Smith says that "Chaplin is funny" is neither true nor false while he is describing a philosophical context; that is to say, he neither asserts nor denies that "Chaplin is funny" is true in that target context.[6] To pass ICD, we must be able to build a real CSA (RCSA) where the alleged context-sensitive expression is either asserted or denied in *every* target context. But Smith's CSA does not pass ICD; therefore, we conclude that the truth predicate does not seem to be context-sensitive in Kölbel's sense.

Let us now consider a more fine-grained CSA.[7]

> Smith is saying that "Chaplin is funny" is true and by that he simply means that Chaplin is funny. There is a false utterance of "the judgment that Chaplin is funny is true", not because Smith thinks that Chaplin is not funny but because "Chaplin is funny" is not true in the same sense as "Alghero is in Sardinia" is true. Indeed, "Alghero is in Sardinia" is true because it is an objective fact of the matter; whereas "Chaplin is funny" is true merely because Smith believes it is true, but he is also aware that taste judgments are not objective.

[6] We assume that if someone says that <p> is neither true nor false, then she neither asserts <p> nor denies <p>.

[7] We thank an anonymous referee for drawing our attention to this CSA.

Which CSA is the above? Is a RCSA or an ICSA? This story is trickier than the previous one, so it requires a bit more endeavor. *Prima facie*, we have two target contexts: one (a) where Smith says that "Chaplin is funny" is true, and one (b) where Smith says that "Chaplin is funny" is not true in the same way as "Alghero is in Sardinia" is true (we can equivalently say that Smith denies that "Chaplin is funny" is true in the same way as "Alghero is in Sardinia" is true).

The second context, (b), can be interpreted in two ways. Under the first interpretation, (b1), Smith asserts that it is true that being in Sardinia (an objective fact of the matter) is not the same as being funny (a taste matter); under the second interpretation, (b2), Smith says that "Chaplin is funny" (a taste judgment) is neither true nor false, and that "Alghero is in Sardinia" (an objective judgment) is true.

Notice that under the second interpretation, (b2), we get three target contexts in total after eliminating the conjunction: first, (a) Smith asserts that "Chaplin is funny" is true; secondly, (b2 E$\wedge$) Smith says that "Chaplin is funny" is neither true nor false; thirdly, (b2 E$\wedge$) Smith asserts that Alghero is in Sardinia. Because Smith neither asserts nor denies that "Chaplin is funny" is true in one conjoint, then the interpretation (b2) yields an ICSA.

Let us now look at the first interpretation (b1). Here we have two contexts: first, (a) Smith asserts that "Chaplin is funny" is true; secondly, (b1) Smith asserts that it is true that being funny is not the same as being in Sardinia. Given the equivalence schema, in (a) Smith simply asserts that Chaplin is funny, and in (b1) he simply asserts that being funny is not the same as being in Sardinia. However, it seems to us that Smith is actually making two assertions, (a) and (b1), about what is funny; not about what is true. At the end of the day, we agree that being funny is not the same as being located in Sardinia, because the former is a subjective property and the latter is an objective property; however, we do not need two meanings of "true" in order to make sense of the distinction between subjective judgments and objective ones. To put it another way, Smith is just saying that being funny is a subjective judgment, whereas being in Sardinia has an objective status.

If our considerations are correct, then the predicate "true" is not context-sensitive in Kölbel's sense. Of course, a possible reply is to blame the theoretical apparatus we have used. This is a fair objection, but our argument aims to prove that "true" is not context-sensitive in Kölbel's sense within Cappelen and Lepore's analysis of context-shifting arguments. Insofar as we adopt their technical apparatus, we can conclude that the ordinary usage of the truth predicate does not seem to be context-sensitive in Kölbel's sense.

5. Conclusions

We have argued that "true" in English is neither ambiguous nor context-sensitive. More precisely, we have claimed that the English truth predicate does not have these properties in the specific way envisaged by Kölbel.

For all we have said, "true" could be ambiguous or context-sensitive in other ways. We allege that the manner in which TRUTH-C is defined, i.e. by means of the notion of faultless disagreement, may be held accountable for the stumbling block to passing the tests. We acknowledge that people have faultless disagreements on taste judgments, but this does not seem to require that truth must be

split into two different concepts; rather it may imply that people tend to be relativistic on matters such as taste judgments. This hypothesis, however, requires further evidence and perhaps even a different methodology. We then encourage the collection of robust empirical data from ordinary speakers in order to shed light on our conjecture.

Moreover, "true" could display an interpretative uncertainty of a different kind than ambiguity or context-sensitivity. In fact, this is something that Kölbel himself concedes when he claims that the ambiguity in question may be a pragmatic phenomenon—one of the two senses of "true" would be conversationally implicated along Gricean lines (Kölbel 2008: 369). And there are other possibilities too. "True" could be vague or semantically indeterminate, as McGee and McLaughlin (1995) have indeed suggested. Our investigation, so far, has left these options open.

References

Cappelen, H. and Lepore, E. 2003, "Context Shifting Arguments", *Philosophical Perspectives*, 17, 25-50.

Cappelen, H. and Lepore, E. 2004, *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*, New Jersey: Wiley-Blackwell.

Cappelen, H. and Hawthorne, J. 2009, *Relativism and Monadic Truth*, Oxford: Oxford University Press.

Fine, K. 1975, "Vagueness, Truth and Logic", *Synthese*, 30, 265-300.

Kölbel, M. 2008, ""True" as Ambiguous", *Philosophy and Phenomenological Research*, 77, 359-84.

Kripke, S. 1979, "A Puzzle About Belief", in Margalit, A. (ed.), *Meaning and Use*, Dordrecht: Reidel, 239-83.

McGee, V. and McLaughlin, B. 1995, "Distinctions Without a Difference", *Southern Journal of Philosophy* (Supplement), 33, 203-51.

Wright, C. 1992, *Truth and Objectivity*, Cambridge (MA): Harvard University Press.

Zwicky, A.M. & Sadock, J.M. 1975, "Ambiguity Tests and How to Fail Them", in Kimball J.P. (ed.), *Syntax and Semantics*, 4, New York: Academic Press, 1-36.

Literature and Practical Knowledge

Pascal Engel

EHESS

Abstract

This article defends literary cognitivism, the view that literature can convey genuine propositional knowledge, in the form of propositions which are (i) true (ii) justified and (iii) have aesthetic value because they convey such knowledge. I reply to familiar objections to this view, and reformulate it as the thesis that literary knowledge is a form practical knowledge that is only derivatively propositional. I attempt to apply some ideas to be found in Stanley's and Williamson's conception of knowing how. Literary knowledge is a kind of practical knowing how of propositions involving demonstrative practical modes of presentation. This conception has often been criticized, rightly, for relying on a notion of knowing how that is too intellectualist. But in the case of literary knowledge, where we never get direct knowledge of experience or practice, and where our knowledge is always mediated by the properties of form and style, this drawback is actually a virtue.

Keywords: Literature, Literary cognitivism, Knowledge, Truth, Knowing how, Practical knowledge, Jason Stanley, Timothy Williamson

1. Introduction: Literary Cognitivism

Although nobody would deny that we learn a lot from reading literary works, as soon as one tries to say more precisely what it means to come to know something from them, the answers become elusive. There is after all a long tradition in literary criticism according to which the aim of literature is to bring us knowledge of the world and of human nature. It is often called "literary humanism". One can find typical expressions of this view in such declarations as Samuel Johnson's:

> The value of every story depends on its being true. A story is a picture either of an individual or of human nature in general; if it be false, it is a picture of nothing (Boswell 1837: 32).

One could find many similar claims in, for instance, writers like Charles Dickens, George Eliot, Henry James, Joseph Conrad, Virginia Woolf, Thomas Mann, and in literary critics like J. Benda (1945), F.R. Leavis (1948) or L.

Argumenta 2, 1 (2016): 55-76
ISSN 2465-2334

DOI 10.14275/2465-2334/20163.ENG

Trilling (1950).[1] What seems to be distinctive of literary humanism is the insistence on the idea that there is such a thing as literary truth and literary knowledge. Many authors, however, are hostile to the view that there can be any kind of truth in literature. Thus Peter Lamarque and Stein Olsen (1989, see also Lamarque 2008) defend what they call a "no-truth" view of literature, according to which what is distinctive of learning from literature is not the fact that it brings us certain kinds of truth, which we would have to know, but the fact that it rests on certain practices, which we would have to imitate. They do not want, however, to deny that the value of literature rests upon its promotion of various humanistic themes, which for them are mostly relative to the social values carried by the times, places and historical contexts from which literary works emerge and that they—directly or otherwise—describe. They insist nevertheless on being called "literary humanists" in this less than universalist sense. It is hard to see, however, how universal values and ideals such as truth, sincerity or justice, which are the traditional ones promoted by literary humanism, can be preserved in this relativistic framework: for how could these values transcend

[1] Some other examples among many such declarations:

"Shakespeare is above all writers, at least above all modern writers, the poet of nature; the poet that holds up to his readers a faithful mirror of manners and of life. His characters are not modified by the customs of particular places, unpractised by the rest of the world; by the peculiarities of studies or professions, which can operate but upon small numbers; or by the accidents of transient fashions or temporary opinions: they are the genuine progeny of common humanity, such as the world will always supply, and observation will always find" (Johnson 1765: 8).

"The picturesque contrasts of character in this play are almost as remarkable as the depth of the passion. The Moor Othello, the gentle Desdemona, the villain Iago, the good-natured Cassio, the fool Roderigo, present a range and variety of character as striking and palpable as that produced by the opposition of costume in a picture. Their distinguishing qualities stand out to the mind's eye, so that even when we are not thinking of their actions or sentiments, the idea of their persons is still as present to us as ever. These characters and the images they stamp upon the mind are the farthest asunder possible, the distance between them is immense: yet the compass of knowledge and invention which the poet has shown in embodying these extreme creations of his genius is only greater than the truth and felicity with which he has identified each character with itself, or blended their different qualities together in the same story" (Hazlitt 1916: 33).

"It is useless to discuss whether the conduct and character of the girl seems natural or unnatural, probable or improbable, right or wrong, IT IS TRUE. Every man who has watched these melancholy shades of life, must know it to be so. From the first introduction of that poor wretch, to her laying her bloodstained head upon the robber's breast, there is not a word exaggerated or over-wrought" (Dickens 1999: lvii).

"A work that aspires, however humbly, to the condition of art should carry its justification in every line. And art itself may be defined as a single-minded attempt to render the highest kind of justice to the visible universe, by bringing to light the truth, manifold and one, underlying its every aspect (Conrad 1914: 12).

"The only reason for the existence of a novel is that it does compete with life. When it ceases to compete as the canvas of the painter competes, it will have arrived at a very strange pass. It is not expected of the picture that it will make itself humble in order to be forgiven; and the analogy between the art of the painter and the art of the novelist is, so far as I am able to see, complete. Their inspiration is the same, their process (allowing for the different quality of the vehicle) is the same, their success is the same. They may learn from each other, they may explain and sustain each other. Their cause is the same, and the honour of one is the honour of another" (James 1884: 46).

the social and historical boundaries to which they are supposed, on this view, to be attached?

The view that I try to defend here shares with the traditional form of literary humanism the idea that there are literary truths and literary knowledge, in the most straightforward and literal senses of the words "truth" and "knowledge". I propose to call this view *literary cognitivism* (LC) and to define it by at least the following three theses:

(i) *Truth condition*: All genuine literary works express general truths about the world and about human nature.
(ii) *Knowledge condition*: Some literary works have a cognitive value in the sense that they express knowable truths and are able to impart them.
(iii) *Aesthetic value condition*: This cognitive value is essential to the aesthetic value of these works.

In what follows, I shall first try to reply to familiar objections against this view. Those who are impressed by these objections have concluded that if knowledge can be conveyed by literary works at all, it cannot be a form of theoretical propositional knowledge, but a form of practical knowledge. This view, however, is very elusive. I reformulate it through an account that tries to reconcile the propositional character of literary knowledge with its practical character.

2. Versions of Literary Cognitivism

It is hard to deal with these issues without begging all sorts of questions about the nature of literature, the nature of truth and the nature of knowledge. First, one thing needs to be said about the scope of "literary". "Literature", as Lamarque and Olsen (1989: 24-25) insist is, unlike "fiction" or "narrative", a normative or evaluative term and not a descriptive one. So "genuine literary works" is question begging if one postulates that genuine literary works do have a cognitive value. Why consider Dante, Shakespeare, Cervantes, Goethe, Conrad or Proust as more worthy of this characterization than Barbara Cartland, J.K. Rowling or Paulo Coehlo? Who knows? Some people consider *The Lord of the Rings* as a much better epic than *The Iliad* and some seem to value *Harry Potter* more highly than Robert Louis Stevenson's or Conan Doyle's novels. The only answer that I can give is that classics are *prima facie* better candidates against sceptics, just as "I have hands" or "The earth was not created yesterday" is a better candidate for something known for certain than "Julius Cesar had a cold when he crossed the Rubicon" or "Julius is the guy who invented the zip". Indeed I cannot prove that J.K. Rowling will stand the test of time and that her novels will lead to the production of as many works of literary criticism as Marcel Proust or Samuel Beckett, but for the time being, I have a better—even if not perfect—guarantee (and plausibly tied to a certain kind of institutional setting, time, practice, etc.) that the latter belongs to literature and not the former. Some may attach to literature values other than the cognitive—in particular emotional value—and for that reason may prefer *Harry Potter* to *La recherche*, but they would equally beg the question in assuming that emotional value is independent of cognitive value. They also would be wrong to think that the *Commedia*, for instance, carries less emotion than, say, *The Da Vinci Code*. LC indeed does not say that any kind of work that aspires to being a piece of literary writing conveys knowledge. It only says that some works at least can achieve this goal.

Second, the plausibility of LC may seem to depend also on literary genres. It is less clear that it applies to lyrical poetry like Byron's *Childe Harold* or to gothic novels like Walpole's *The Castle of Otranto* than to historical narratives like Gibbon's *Decline and Fall* or to pieces of literary journalism like Truman Capote's *In Cold Blood*. Everything depends upon the notions of knowledge and truth that are question. The notion of truth is used in many diverse senses according to the literary genre to which it is applied. There is a notion of truth that is said to be proper to poetry alone. According to the romantic conception of literature in particular, poetry is the vehicle of a kind of transcendent or essential truth that can be reached through some sort of mystical intuition or revelation. On this view, Novalis' *Hymnen an die Nach* or Mallarmé's *Poésies* contain more, and better, truths than any piece of prose. Zola, for his own part, meant to write "experimental novels" and to elevate literature to an almost scientific description of characters in their biological and social setting. Other writers, such as Paul Bourget, intended to contribute to psychological science. But do we want to restrict our notion of literary cognitivism to these specific enterprises? Postmodernism and literary formalism have made us familiar with the idea that literature describes worlds where no notion of truth or reference whatsoever can apply, or, if they do, such a notion rests upon its own and specific kind of truth, "truth-in-fiction", or "novelistic truth", which has nothing in common with ordinary, garden-variety, truth. Neither do we want to tie literary cognitivism to the idea that a number of true statements often occur, among others, in fictional narratives, such as the first sentence of the *Chartreuse de Parme*, "On the 15th of May, 1796, General Bonaparte marched into the city of Milan", or the first sentence of *La recherche du temps perdu*, "Longtemps je suis couché de bonne heure", which is probably true (but of whom? That is a more difficult question). Clearly if LC is to make sense, one has to say upon what kind of notion of truth it rests.

So there are quite a number of versions of literary cognitivism, depending on the notions of truth and knowledge that one is committed to. Let us start with what appears to be the strongest condition, the knowledge condition (ii).

The notion of "cognitive value" of a literary work of art is vague indeed. Very often it means that fictions, narratives or other kinds of literary works are apt to lead readers to infer, or perhaps to consider, through some form of imaginative understanding, a number of beliefs. Many of these beliefs, on whatever subjects—say moral beliefs, or beliefs about human psychology—may be true, hence have a cognitive value. Similarly the work of imagination can enhance our cognitive powers. It is not clear, however, that literary works can do so by leading us to form *new* beliefs, let alone new true beliefs. If on the basis of reading Kipling's *The Man who would be King*, I form the belief that *Kafiristan* is a country located in the mountains north of Afghanistan, do I form a belief that is true? No. But I can form a belief that there is such a country, or I can try to identify Tajikistan under the fictional name of Kafiristan. But not all beliefs are so empirical. If on the basis of reading *Effi Briest*, I form the belief that women are oppressed in marriage, it may not be something that *I learn* from the novel or from that novel only. It is not clear that I form a new belief, and many have argued that the only kind of knowledge that one can find in literary works is based on previously acquired beliefs or knowledge, hence is more a form of recognition than a form of cognition (Stolnitz 1992). Moreover how could I identify the truth that marriage is oppressive to women in Fontane's novel if I had no knowledge of social relationships in nineteenth century Germany? And is it the

same knowledge as the one I form through reading *Middlemarch*? The idea is present in a number of views according to which literature brings a cognitive strengthening of what we already know, but not *new* knowledge.[2] The same idea is much present in a simple form in contemporary analyses of fiction. Thus David Lewis famously remarks about fiction:

> Most of us are content to read a fiction against a background of well-known fact, "reading into" the fiction content that is not there explicitly but that comes jointly from the explicit content and the factual background (Lewis 1975: 268).

And Lewis suggests that fictions are pieces of counterfactual reasoning:

> Reasoning about truth in fiction is like counterfactual reasoning: we make a supposition contrary to fact and […] we depart from actuality as far as we must to reach a possible world where the counterfactual supposition comes true (*ibid*: 269),

or, in terms of beliefs, we form hypothetically a given belief and see, on the assumption that it is true (or probable), whether a consequent belief is true (probable) or compatible with the first.[3] This view is often associated to the familiar idea that fictional narratives involve thought experiments whereby we are invited to ask ourselves the question: "what if….?" and to try to imagine what would be the case if one entertained the supposition described in the antecedent of a conditional. Indeed if thought experiments can sometimes contribute to the formation of scientific or philosophical theories, it is tempting to suggest that fictional narratives, in so far as they involve thought experiments, can contribute to the formation of knowledge. But how this is achieved is a moot question.[4]

It is one thing for a piece of literary work to contribute to the formation of beliefs, including true ones, or even to contribute in some way to the formation of knowledge—for instance by suggesting important hypotheses—and it is another thing to be a genuine *source* of knowledge, and quite another thing again to *impart*, or to *transmit* knowledge in the way ordinary learning is supposed to do. It is yet another thing to be accidentally a vehicle of a true belief or of knowledge—as for instance when I overhear someone saying something that turns out to be true and I come to believe it—and to be *essentially* a source of knowledge or to constitute a form of knowledge. Presumably many of those who claim that they "learn" from literature have only the former—weak—sense in mind, which is insufficient for granting that literature has cognitive value.

If literary cognitivism is to have some bite—if it is not to be trivial or empty—the notion of knowledge involved in thesis (ii) had better be robust. In other words, it had better coincide with the ordinary common sense notion of knowledge, which involves the condition that a belief be not only true, but also justified or warranted.[5] Now the ordinary notion of knowledge applies, first and

[2] Carroll 1998, Gibson 2009. The idea is indeed familiar from philosophical hermeneutics.

[3] Anyone familiar with the literature on conditionals will recognize the so-called Ramsey's test on conditionals.

[4] On this point, see Lombardo 2012, Engel 2012, Ch. 5.

[5] *Pace* some experimental psychologists' claim to the contrary. Indeed that knowledge can be so *defined* is another matter. Currie to appear studies various ways in which we can get

foremost, to *propositional* knowledge, which involves a relation to propositions (whether expressed by sentences or not) that can be true or false. Knowledge in the ordinary sense, being factive—to know that P entails that P is true. Let us call this view the *strong* form of literary cognitivism.

Before examining this point, we must say something about the notion of truth involved. The strong form of literary cognitivism has to go with a notion of truth which is robust enough to be applied to propositions that can be true and known, lest we beg the question by implicitly accepting a weaker notion of knowledge (say, as cognitive improvement of the reader) or a weaker notion of truth. Some notions of truth are so weak or so shallow that they apply to many forms of discourse, from ethics to fiction, from aesthetic truths to scientific truths to commonsense truths or to comic truths. If, for instance, we accept that there are fictional truths, moral truths, legal truths, mathematical truths, and so on and so forth, and as many truths as there are possible objects of discourse, the concept of truth becomes completely trivial, and there is no way to distinguish the domains where it applies from those where it does not apply. Deflationists about truth, who say that "true" applies wherever we can apply the equivalence schema "'P' is true if and only if P", will agree and will welcome that conclusion.[6] They tell us that if we can talk of fictional truths at all, then indeed we can frame such equivalences as "'Sherlock Holmes lives in Baker Street' is true if and only if Sherlock Holmes lives in Baker Street". Such sentences tell us nothing about whether there can be fictional truths if we assume from the start that they are true *in fiction*, and similarly for all kinds of truths. Or if we are ready to talk about comic truths, that "'Charlie Chaplin is funnier than Groucho Marx" is true if and only if Charlie Chaplin is funnier than Groucho Marx". Such equivalences would mean that our schema is relative to a domain, or to a framework, or to a kind of discourse, and it would follow that there are as many kinds of truth as kinds of discourse.[7] If we do not want to trivialize truth in this way, we must accept that the concept of truth carries more weight and does not reduce to the innocuous equivalence between "'P'" and "it is true that P" (in such and such domain, for such and such discourse). We must accept the idea, implicit in the schema "'P' is true if and only if P" or "the proposition that P is true if and only if P", that the right-hand side of these schemata tells us something about what has to be the case *in the world*—the actual world, and not in some fictional or legal, or moral, etc. world—for the sentence or proposition of the left-hand side to be true. In other words one has to accept that truth involves at least the idea that for a proposition to be true there must be something in virtue of which it is true, hence that truth involves some sort of relation (of correspondence or other) between what a statement describes and the way things are. We need not spell out what kind of concept of truth this implies, but we must at least accept that if a statement is true, it obeys a certain minimal set of platitudes: the equivalence principle "'P' is true if and only if P", the fact that truth is the aim of assertion and belief, that it is objective and independent of our investigations at least in the sense that a statement can be justified but not true, and

beliefs from fiction. But he is skeptical about our ability to get genuine knowledge from them.

[6] E.g. Horwich 1991, Rorty 1991.

[7] This is what is sometimes called "alethic pluralism". For familiar objections to it, see Engel 2009, Lynch 2009.

that statements are true if and only if they describe things the way they are.[8] In other words, and in so far as these platitudes entail that the notion of truth involved is the ordinary one, the notion of truth has to be robust. The same argument applies to the notion of knowledge. We can have a very shallow notion of knowledge, according to which wherever there is a justified true belief of any kind, there is knowledge. Thus if we are in some sense justified (say, by the number of laughs) to say that it is true that Charlie Chaplin is funnier than Groucho Marx, then we can say that we *know* that Charlie Chaplin is funnier than Groucho Marx. Knowledge is usually associated with standards of justification, which can be high or low. But if we accept that any standard of justification, however low, counts for a true belief to be knowledge, then the notion of knowledge trivially applies virtually to every truth, hence to literary truth, provided that there is such a thing.

A belief, however, to be apt to be knowledge, must at least be true. So even when the notion of truth is reduced to these platitudes it seems always to be *too* robust to be applied to literary discourse, and in particular to fictional narratives. How can one claim that literature brings cognitive content which can be knowledge *in virtue of the fact that it is most of the time intrinsically fictional and does not aim at truth*? Fiction is by definition non-veridical and non-referential. So how can it be knowledge *in virtue of its suspension of reference and truth*? Putting the bar of cognitive value so high seems to raise immediately the threat of scepticism with respect to literary cognitivism.

LC certainly cannot be true in the sense that all sentences of a literary work of art are supposed to be true, unless one means to restrict the thesis only to works with truth-telling objectives, such as historical narratives, reports, journalism, or literary and philosophical essays, and if we suppose (which is not obvious) that these kinds of writings actually aim only at expressing truth and are all truth-apt. Nor, as we have seen, is LC the thesis that the poets have to be expelled from the city, and that literature has to become a sort of science.[9] So the propositional truths that feature in a literary work cannot be those expressed by its very sentences, even when they happen to be true as a matter of fact (as when one reads in a novel: "Hitler invaded Poland in September 1939"). They must be general truths that the reader, whether or not they are intended by the writer, believes that the literary work in some sense expresses or contains, and that he can extract from it. This is the view which is often called the "message" conception of literature. Lamarque and Olsen call it "The Propositional Theory of Literary Truth":

> *The literary work contains or implies general thematic statements about the world which the reader as part of an appreciation of the work has to assess as true or false*. The theory

[8] For this approach to truth, see Wright 1993, Lynch 2009. Among the platitudes the correspondence platitude is prominent. Approaches to truth through the platitudes are often associated to a pluralistic view of truth (see Lynch 2009), according to which truth may conform to the platitudes in general, but can be "realized" in different ways depending on the domain. The thought then would be that, unlike, say, in physics or mathematics, truth in literature could be a form of coherence rather than a form of correspondence. This is not the view that I intend to defend, when I say that our concept of truth in literature has to be robust. I take it as a correspondence truth here too. Thanks to Michael Lynch for having pointed this out to me.

[9] As Benda (1945) sometimes suggests. See Engel 2012.

> presents two claims. First, a literary work implies propositions which can be construed as general propositions about the world. Second, these propositions are to be construed as involved in true or false claims about the world. In the terminology of *theme* and *thesis* the theory would be that a literary work develops not only a theme but also a thesis and that part of the appreciation of a literary work as a work of art is an assessment of the truth-value of this thesis (Lamarque and Olsen 1989: 325).

For instance, in George Eliot's *Middlemarch,* the Lydgate story, according to Eliot herself, shows that "the best human hopes and aspirations are always thwarted by forces beyond human control". Such general, indeed rather banal, truths are very often taken to be psychological laws of human nature. They very often form the content of the moral knowledge that is, on many classical and contemporary views, conveyed by literature.

The problem with the Propositional theory, as Lamarque and Olsen argue, is that thematic statements, explicit as well as implicit, can be assigned significance and thus be understood without being construed as *asserted* (1989: 328). This raises three kinds of problems.

a) The first may be called the *interpretation* problem: how are the themes or general truths expressed by a literary text, and if they are implicit, how are they accessed by the reader who is supposed to retrieve or extract them? Presumably through a process of interpretation, but how does it work?

b) The second may be called the problem of *triviality*, just indicated about Eliot's declaration about Lydgate. Whether or not these general truths are explicit, they are bound to be in many ways humdrum and trivial, as most take home messages which one can get from famous novels, such as: "All happy families are alike; each unhappy family is unhappy in its own way", "It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife". Or even more platitudinously: "provincial life is boring", "human nature is bad", "life is often made of difficult choices", "It is no good to stay in bed all day"; etc. If it is the role of literature to bring to us such trivialities, or at best to repeat proverbs from the wisdom of nations, its knowledge content is very poor indeed.

c) The third objection is related to the previous ones. Suppose that the message or theme is complex and informative—rather than dull or trivial—and that it brings us some deep and complex truths about human nature. If it does, in what sense can it do so independently of the form and style of the literary work? A common version of this objection can be called the problem of *paraphrase*: if a literary work is to deliver a specific knowledge content, this content must be apt to being expressed in other, presumably more accessible, terms. But if literary works and fictions were paraphrasable into plain truth-aimed prose, the literary form—what gives to the work its aesthetic value—would be lost. The style, the writing—everything that makes for the value of a literary work—will be inessential, and only the content will be attended to. Not only, as most literary critics have argued, is it very dubious that one can so separate form and content in a literary work, but even if we could do so, it would be a complete misrepresentation of the nature of literary work. If, more plausibly, the general truths that are supposed to be transferred by literary works are, so to say, embedded within the narrative, how can the knowledge of human nature that they are meant to confer be separated off from its incarnation in the story? If it is not incarnated, our

interpretation is false. If it is, the separation of form from content makes it useless: why would writers care to write novels or stories if their take-home message can be paraphrased in other terms? And why would readers care to read these? George Eliot, who was a stern defender of literary cognitivism, states this dilemma quite well:

> Suppose a language which had no uncertainty, no whims of idiom, no cumbrous form, no fitful shimmer of many hued significance, no hoary archaisms "familiar with forgotten years"—a patent deodorized and non resonant language which effects the purpose of communication as perfectly and rapidly as algebraic signs. Your language may be perfect medium of expression to science, but it will never express life, which is a great deal more than science (Eliot 1883: 287-88).

Eliot here formulates clearly the principle (iii) of LC: it is *in virtue of its being fiction*, that is, in virtue of its *stylistic form* and as the bearer of *aesthetic value* that a literary work carries cognitive value. Eliot, however, does not tell us *how* a literary work can "express life" and thus be a source of knowledge.

Although they are most common, it is not clear that these objections from interpretation, triviality and paraphrase are actually damaging for LC. For literary works, if they are meant to have cognitive value and to carry knowledge, do not wear, so to say, their content on their sleeves. Readers and literary critics have to extract it. If this content consists in general themes about human nature or the world, it might be universal to the point of being trivial. Actually the rich tradition of *ethopeia,* the description of human characters, from Aristotle's *Ethics* and Theophrastus' *Characters*, to the French and British moralists and to Jane Austen and George Eliot, rests upon the description of stable and well-known features of human nature.[10] Ambition, jealousy, greed, pride are everywhere the same, in spite of the variety of situations and of people who exemplify them. What we learn from the repetition of these features may be quite dull. When in *Middlemarch* we read the Lydgate story, do we really learn that "the best human hopes and aspirations are always thwarted by forces beyond human control"? We do not actually learn this from the story, as a lesson, or a message conveyed by it and which could be conveyed by other means (for instance by a philosophical treatise). What we learn in reading novels, or comedies, that bear on these general features of human nature is *how to recognize* the theme in the story, and *how* it is exemplified by these characters in this particular setting. We learn, in other words, that *this is the way* in which human hopes are thwarted by forces beyond human control. The literary work *shows*, *expresses* or *displays* the theme or maxim in question. It does not articulate it, or explain it explicitly and propositionally. For this reason a number of writers have suggested that if literary works impart or transmit knowledge, this cannot be a piece of theoretical knowledge, expressed by propositions, but a form of *practical* knowledge.

To summarise the points advanced in this section, we can represent them in the form of two dilemmas.

The first dilemma involves a choice between a weak and a strong form of literary cognitivism:

(A) If on the one hand, one opts for a weak form of literary cognitivism, according to which there is literary knowledge, but only in the sense of

[10] On this tradition see in particular Van Delft 2012, Carnevali 2010.

gaining, through literary work a capacity to form new beliefs, including true ones, perhaps through some form of inference or some sort of activity of the imagination, or a capacity to reflect on the beliefs that we already have, then it is dubious that these capacities give us a kind of knowledge, in the sense of a justified true belief. At best we enlarge our cognitive capacities. But this does not amount to gaining actual *knowledge*.

(B) If, on the other hand, we opt for a strong form of knowledge (as justified true belief), then we are led to scepticism about literary fiction knowledge: for how can a work of fiction aim at producing knowledge of the ordinary sort and have an aesthetic value?

The second dilemma is between a propositional and a practical view of literary knowledge:

1. If there is literary knowledge, it has to be propositional.
2. But if literary knowledge is propositional, literature cannot provide such knowledge.
3. So either there is no literary knowledge or literary knowledge is not propositional.

To this second dilemma I now turn.[11]

3. Literature and Practical Knowledge

In the light of these difficulties for the propositional theory of literary knowledge, many writers who nevertheless insist that there is such a thing as literary knowledge have argued that the kind of knowledge that is brought by literary works cannot be of the same kind as the knowledge that consists in information, a body of truths, of a propositional or of a theoretical sort. They have argued that literature is not a way of doing philosophy, ethics, history, journalism or science by other means, but that it does not follow that it does not convey a form of knowledge, yet one that does not consist in the expression of beliefs and of truths. They have held that the knowledge in question is practical, a form of *knowing how* rather than a form of knowing that. These claims have taken various forms, all supposed to rest on a strong contrast between the cognitive benefits of learning from literature and its practical effects. Three kinds of claims in particular have been made.

a) First it has been said that literary works involve not only the exercise of imagination, often under the form of a kind of projection or mental simulation, or a form of empathy involving a capacity of readers to share emotions and feelings with the characters depicted in fiction (Walton 1986, Currie 1990, 2010), but also the capacity to *enlarge* our imagination through fictions and narratives. This kind of empathy can involve a form of imagination that is voluntary, creative and recreative (Currie and Ravenscroft 2008). But it can also involve a form of direct participation, a capacity of immediate identification with the characters

[11] Gaskin (2013) defends a version of literary humanism or cognitivism which is distinct from both Lamarque and Olsen no truth view and the practicalist version presented here. Gaskin holds that literature does convey truth and knowledge about the world. But this thesis is defended on the assumption that the world is itself propositionally structured, which he calls linguistic idealism. A number of Gaskin's arguments could be adapted here, but I do not want to defend literary cognitivism at this cost.

and feelings described in stories, which has often been called "learning what it is like" to be a certain sort of person in a certain sort of situation.

b) It has been said also that literary fictions give us a capacity to see and to understand human situations, by giving us some new vision of life (Murdoch 1997). Thus Putnam writes about Céline's *Voyage au bout de la nuit* that in reading this novel he does not "learn that love does not exist, that all beings are hateful and hating […] What I learn is to see the world as it looks to the eyes of someone who is sure that that hypothesis is correct" (Putnam 1978: 89). Thus Jacques Bouveresse writes:

> If one can derive any knowledge, in particular knowledge of human beings from our acquaintance with a literary work, it seems to consist only in some sort of practical knowledge. What teaching and learning look like has nothing to do with the communication of a theory, including the kind of theory that the author may be able to develop by himself. It's only because literature is probably the most appropriate means to express, without falsifying them, the indeterminacy and the complexity of moral life that it can learn to us something essential in this domain. To take up Wittgenstein's phrase, it can help us to watch and to see many more things than what ordinary life would allow us to watch and to see—at the very moment when we are tempted, a bit too early and too fast, to think (Bouveresse 2008: 54).

c) It has been argued that literary works can make us improve our capacities or give us some skills through some sort of training or education akin to a drill but also to some kind of formative process or *Bildung*:

> While it is often assumed that fictions must be informative or morally improving in order to be of any real benefit to us, certain texts defy this assumption by functioning as training grounds for the capacities: in engaging with them, we stand to become not more knowledgeable or more virtuous but more skilled, whether at rational thinking, at maintaining necessary illusions, at achieving tranquillity of mind, or even at religious faith. Instead of offering us propositional knowledge, these texts yield know-how; rather than attempting to instruct by means of their content, they home capacities by means of their form; far from seducing with the promise of instantaneous transformation, they recognize, with Aristotle, that change is a matter of sustained and patient practice […] Increased agility makes us better at doing what the text expects of us, which in turn leads to still greater agility not just as reader but, more generally, as liver of a life. Thus rather than providing knowledge per se—whether propositional knowledge, sensory knowledge, knowledge by acquaintance, or knowledge by revelation—what such texts give us is know-how; rather than offering us a new set of beliefs, what they equip us with are skills; rather than teaching, what they do is train. They are not informative, that is, but formative. They present themselves as spiritual exercises (whether sacred or profane), spaces for prolonged and active encounters which serve, over time, to hone our abilities and thus, in the end, to help us become who we are (Landy 2012: 12).

Such claims, which are by no means equivalent, are in many ways puzzling. In the first place, if the exercises of imagination involved in claim a) are supposed to be cases of *knowing how*, and if *knowing how* is understood, along the classical lines of Ryle (1946), as distinct from and irreducible to *knowing that*, it is not clear that they involve contents completely alien to propositional and con-

ceptual understanding. Typically, the ambiguous phrase "what it is like" may refer to an experience, of a qualitative sort, accessed in the first person—as the experience of a colour or of a taste—and which is indeed—at least on most views of *qualia*—non-propositional and non-conceptual. But it may also refer to a habit or a disposition to have this kind of experience, which may also involve capacities of recognition that are not purely experiential. It is clear that someone who experiences what it is like to eat a pineapple for the first time learns from this experience, but can we transpose this claim to a reader of a novel who "experiences", through his reading of, say, Solzhenitsyn's *A day in the life of Ivan Denisovich* what it is like to be a prisoner in the Gulag? By definition the experience is conveyed to the reader in a way in which the taste of the pineapple cannot be conveyed. It has to be the description of an experience, not the experience itself. If it is conveyed to the reader, it is through an indirect description, under the form of a testimony, not through some direct experience. Some philosophers hold that "knowing what it is like" can be reduced to a form of knowing how (Lewis 1990, Nemirow 1990). But it does not follow that this can be transposed to a literary "what it is like". In the second place, if the kind of knowing-how is a form of knowing how human beings live, or knowledge of a form of life, it is bound to be quite elusive, so elusive indeed that it can be doubted that it is a form of knowledge at all. What is knowledge that "expresses the complexity of life" with all its "indeterminacy", and that learns "to watch and to see"? What would be a knowing how about *life* in general? The claim that it is life as a whole that we learn through literary works actually comes close to the view that it conveys no knowledge at all, and seems rather to express a form of skepticism about the cognitive value of literature. The capacities and skills which Landy claims to constitute the literary knowing-how seem to be more specific, but it is not clear what kinds of skills and know-hows are conveyed by literary texts. They cannot be particular pieces of physical or technical know-hows, such as knowing how to ride a bicycle, sailing, playing a musical instrument or using a tool. I cannot learn how to sail by reading Conrad's novels, or how to commit a murder by reading detective stories (although I can learn quite a number of facts about these activities). It seems clear, as Landy says, that they have more to do with the exercise of certain intellectual capacities and activities, such as imagining, thinking, reasoning, or with the capacity to recognize certain emotions or feelings and to be able to transfer these from the characters and situations to one's own case. But the further these are from the exercise of a physical aptitude, the harder it is to believe that they do not involve any kind of propositional knowledge. For instance learning skills in "rational thinking" or "maintaining necessary illusions" can hardly pass for a piece of knowing-how involving no propositional or conceptual competence. The same holds for the exercise of imagination referred to in claim (a): some imaginings are better described as exercises of counterfactual reasoning (in Lewis's sense quoted above) than as cases of direct empathy. Ryle (1949) famously argued that mental states, such as desires, beliefs or intentions are better conceived as dispositions to act. But he also argued that no mental state can be defined by a single disposition, and that it must consist in a, perhaps open-ended, set of disposition. If the same holds of literary practical knowledge, it is very likely that the number of dispositions that are manifested by it is equally diverse and open-ended.

This is all the more true about a fourth d) conception of practical knowledge, invoked by Nussbaum in particular, and modelled on Aristotle's

conception of practical reasoning, according to which this knowledge is the product of a practical syllogism, with a particular premise, a general law and where the conclusion is an action. Such reasoning is not a form of non-propositional knowing-how. On the contrary it involves the exercise of practical judgement, and of deliberation on the basis of reasons.

In spite of the fact that the claims (a)-(d) are rather distinct, let us call this set of views *practicalism about literary knowledge*. Many theorists of literature have found this view much more plausible than literary cognitivism. But it seems that it cannot be reconciled with propositionalism. Indeed the following set of claims, made by those who take literature to impart practical knowledge *and* to impart moral *truths* seems inconsistent:

(i) Literary works aim at producing general truths about human life.
(ii) Literature conveys some kind of moral knowledge.
(ii) This kind of knowledge is practical knowledge.

How can the moral knowledge imparted by literature be both a set of truths about human life and a kind of practical knowledge?[12] If we want to make them consistent, we have either to renounce propositionalism or to reject practicalism. These claims, are, however, consistent, if one rejects the Rylean view of practical knowledge. Here I want to defend practicalism, but I want to argue that it is not incompatible with literary cognitivism. Literary knowledge is not a kind of knowledge that is *directly* theoretical and propositional: literary works and fiction do not aim at producing statements, whether explicit or not, that have a truth value and that are justified by reasons. It does not follow, however, that LC is wrong. Literary works can impart knowledge which is *both practical and propositional*.

4. Intellectualism about Literary Knowledge

I have suggested that the practicalist's claim that literary knowledge consists in a form of knowing-how encounters immediate difficulties if one insists, in Rylean fashion, upon drawing a sharp distinction between propositional and theoretical knowledge on the one hand, and practical knowledge on the other hand. But the boundaries between the two kinds of knowledge might not be so clear, as Stanley and Williamson (2001) have argued. I shall not here deal with their much discussed arguments in detail, and content myself with summarizing them.[13] First they argue that Ryle's regress argument, according to which practical knowledge cannot rest upon the prior contemplation of a proposition is misguided. Second, a number of typical cases of *knowing how*, such as knowing how to fix a car, to find one's way in a city, or to sail a boat, may depend on previously acquired propositional knowledge. Third, and this their main argument, they claim that in English and in German at least the *wh*-constructions that serve to express knowing-how have a deep syntax or logical form which does not differ from that of constructions that serve to express knowing-that. In other words *knowing-how* constructions should not be parsed on the scheme of verb-phrase + infinitive, but on the scheme of an indirect or "embedded" question, wherein X

[12] Bouveresse (2008) makes this puzzling set of claims (I)-(III), which are inconsistent. See Engel 2012, Ch. 5.
[13] For discussions of these arguments, see in particular Rumfit 2012, Benson and Moffett 2012, Hornsby 2012, Wiggins 2012.

is said to know the answer (or an answer) to the direct question: 'How is one to ϕ?' So sentences such as

(1) Hannah knows how to ride a bicycle

should be parsed as belonging to a family of sentences of the form:

(a) Hannah knows whom to call for help in a fire
(b) Hannah knows why to vote for Gore

where "knows" has a propositional, hence truth-evaluable, complement. Stanley and Williamson take this as showing that there are strong grounds for reducing knowing *how* to knowing *that*. They give an analysis of knowing-how in terms of "practical modes of presentation", which are a variety of demonstrative Fregean senses, "ways" of doing this or that in a contextual setting. To know how to ϕ is to know a *way* of ϕ-ing, which is a practical mode of presentation for ϕ-ing:

> X knows how to ϕ iff for some [contextually relevant] way *w* which is a way for X to ϕ, there is a practical mode of presentation *m*, such that X knows under *m* that *w* is a way for her to ϕ (Stanley and Williamson 2000: 428).

In this sense for them the upshot of their analysis is that:

> The analysis is thoroughly intellectualist; knowing how to F is a matter of having propositional knowledge. Like *all* knowledge attributions, intuitive judgments about the truth or falsity of such judgments are sensitive not just to the components of the proposition putatively known, but also to the way in which the subject thinks of them (Stanley 2011: 202).

So on this view, there are strong grounds for reducing practical knowledge to propositional knowledge.

There are, however, strong objections to Stanley's and Williamson' s view.[14] On the linguistic side, if their view is supposed to rest, at least in part, on the semantics of natural language constructions, Rumfit's (2003) objection carries weight. Rumfit notes that the grounds for reducing *know how* constructions to *know that* and *know wh-* constructions are slight: French and Russian construct know how with infinitives rather than with *wh-* sentences but French allows both:

(i) Jean sait préparer la tarte tatin (*Jean knows how to prepare tarte tatin*).
(ii) Jean sait préparer la tarte tatin avec un four micro-ondes (*Jean knows how to prepare tarte tatin with a micro-wave*).

The infinitive construction is not the same as the *wh-* interrogative one, although it is no less natural in French and Russian. This threatens Stanley's and Williamson's claim that the interrogative is the deep structure of knowing-how ascriptions.[15]

A second objection is that the linguistic expressions of knowing *how*, even when they turn out to be expressed with propositions, do not necessarily reflect the nature of the practical knowledge involved. These may well be, in Ryle's

[14] A number of these objections can be found in the essays in Bengson and Moffett 2012.
[15] I shall not here consider the twists and turns of this debate, in Stanley 2011 and 2011a.

phrase, the "stepchildren" of practical knowledge. This idea is voiced by David Wiggins:

> A ship's pilot who is retained by the maritime authorities to bring large ships safely to anchor in an awkward or difficult harbour can tell us, on the basis of his competence and experience, that when the wind is from the north and the tide is running out, the best thing to do is to steer straight for such-and-such a church tower until one is well past a certain bend in the channel. Almost anyone can come to possess that propositional knowledge but the information they get in this way will probably rest indispensably upon the experience and practical knowledge of a handful of people with a different kind of knowledge, namely practical or (as I shall suggest we say) agential knowledge. The propositional knowledge is the stepchild (if I may borrow Ryle's own metaphor—see his 1945, 6, par. 25) of the pilot's practical or agential knowledge (Wiggins 2012: 109).

In other words, we may express the practical knowledge in propositional form, although it is in its nature a much more complex phenomenon, constituted by a set of abilities which need not be answers to specific propositionally expressed questions, which in some sense condense these abilities, without our being able to *read off* this practical knowledge from its linguistic expression. A related remark is that knowing how to ϕ may not amount to knowing one way of ϕ-ing, but a set of ways of ϕ-ing. It is not clear which one is referred to when one says that X knows how to ϕ.

A third objection, also related to the foregoing is that the practical modes of presentation do not have the proper level of generality. A person can successfully manage to type a word, say "Afghanistan", although be just lucky at getting the proper result, whereas another can do the very same thing successfully, but out of a skilled practice. They both know "how to type 'Afghanistan'" and they know ways of doing just this, but their ways or practical modes of presentation are very different. Alternatively a person can know a way, say, to prune roses, but exercise this practical mode of presentation in different ways. What is known by a person who knows how to ϕ needs to be somehow *generic*, and that is why it cannot be captured by citing particular instances of the person's ϕ-ing (Hornsby 2012).

There are a number of other objections, with which I am not going to deal here, that show that Stanley and Williamson's intellectualist conception of practical knowledge is not likely to succeed.[16] The upshot of these objections is that knowing a practical mode of presentation for ϕ-ing (a way to ride a bike, to cook a meal, to whistle with one's fingers, to do a French kiss, etc.) is not knowing how to ϕ. It is only a demonstrative *description* of a knowing how. Jane may know *a way to ride a bike* without knowing *how to ride a bike*. A demonstrative description of ϕ-ing may be a good summary, or a good guide for a successful ability or a knowing how to ϕ, but it need not be a piece of knowledge of how to ϕ. What the practical modes of presentation give us are a best partial and *indirect* descriptions of knowing-how. But knowing-how does not consist in these modes of presentation, and these do not impart practical knowledge. By "indirect" I

[16] Another interesting objection is that practical knowledge, unlike propositional knowledge, cannot be Gettierised (mentioned by Stanley and Williamson 2001: 425; see Poston 2009). But it is not clear that it succeeds.

mean that knowing a way to ϕ is, to take up Hornsby's phrase, knowing a particular instance of ϕ without knowing how to ϕ *generically*.

This may seem to be bad news for the view, that I am here trying to put forward, which aims at conjoining practicalism about literary knowledge and intellectualism. But it is not. The fact that intellectualism about practical knowledge does not give us a full *reduction* of this knowledge to pieces of propositional knowledge, but that practical modes presentation can serve at best to provide *indirect descriptions* of pieces of practical knowledge, can be used as an argument for intellectualism about practical literary knowledge. How?

When one reads a good, meaningful, and crafted literary work, and interprets it correctly, one does not learn something which can be expressed propositionally, in the form of facts, particular or general. For instance one does not learn, when reading *La recherche du temps perdu*, or *Lord Jim,* facts or laws about human psychology or about human nature. Rather, what one learns is a form of practical knowledge. Which? Not the skills or abilities which are described in the work. For instance one does not learn how to behave within a certain kind of society when one reads *La recherche*, or how to navigate in the Indian Ocean. The novels do not impart *directly* such knowledge. One learns, rather, an indirect description of *what it is like* to be a character of a certain type, within a certain social and historical setting. Contrary to what is sometimes said, one does not learn a direct *what it is like* by reading and understanding a literary work of art. Thus when one reads Kafka's *Metamorphosis* one does not learn what it is like to be, or to become a beetle. When one reads Primo Levi's *If This is a Man*, or Shalamov's *Kolyma Tales* one does not learn what it is like to be arrested and sent to a concentration camp or to the Gulag, but one learns what *it would be like*. One learns about a possible kind of life, and how to recognize this possible kind of life. One does do acquire a particular skill in the way one could, for instance, learn how to prune the roses or ride a bike. One learns a more complex skill, that is to say how to recognize a certain kind of situation or character. Thus when I read Balzac's *Les illusions perdues*, I do not learn a set of facts about starting a literary career in the French world of journalism and salons of the first decades of the nineteenth century, but I learn how to recognize a certain kind of character and of behaviours. This is indeed a skill, which an experienced reader is able to master, but the literary mode of imparting that skill is not direct, as it could be when, learning how to fish trout from a wizened and experienced fisherman, he could show me how to do it by doing the gesture and say to me: "That's how to do it". Some gestures are indeed abbreviations for more complex things that are hard to spell out in words, as when one presents one's fist angrily against someone's face and says: "How do you like that?" This may be a "practical mode of presentation" or "way" of performing an insult, which could be spelled out in so many words.

Literary fictions and narratives are ways of showing certain features of reality without necessarily describing these as parts of reality. What a narrative shows is an aspect of reality, which the author shows *through a practical mode of presentation*:

This is how hate (jealousy, sloth) operates
This is what it is like (to live this kind of life)
This is how it feels.

If one accepts this characterization of literary knowledge as imparting to us various modes of presentation that are propositional, but that describe, indirectly, pieces of practical knowledge, there is no incompatibility between practicalism and propositionalism. We can take the claims (i)-(iii) at the end of section 3 above as a consistent triad.

Literary works never give us a direct practical knowledge, in the way a fencing master can teach you how to fence, a music teacher how to play an instrument, or a businessman how to make a deal.[17] But they can impart this kind of knowledge indirectly, by giving you guidelines. If we sort out kinds of knowledge, we can distinguish at least direct knowledge—or forms of knowledge from acquaintance—of *qualitative* experiences, or *personal knowledge*—what one learns from one's own case through various experiences acquired over time, from what one learns from *testimony*, through the transmission of first-hand knowledge by second-hand knowledge. Literary knowledge is never of the first two kinds, always of the third kind. It is a form of testimonial knowledge. Testimonial knowledge is most of the time propositional, except in those rare cases where one can acquire a way of ϕ-ing by ostension. But literary knowledge is never knowledge by acquaintance nor knowledge by ostension. It is a form of deferential knowledge.[18] By representing reality through a certain aspect, literary narratives *defer* to the reader the knowledge by ostension ("This is the way to be jealous, angry, slothful, ignominious, etc."). But they do not impart it directly. So they do not convey *directly* the know-how. One can transmit a piece of knowing-how to someone who does not have it through a practical mode of presentation: this is how to ϕ. Taken together, the practical modes of presentation displayed by literary narratives do not give us any genuine know-how (one does not know how to navigate in the Southern seas by reading Conrad, one does not learn how to be slothful by reading *Oblomov*, how to become a prostitute by reading *Moll Flanders*, how to poison by reading Dickens' "Hunted Down", how to become virtuous through reading *The Mill on the Floss.* But one learns *a way* of sailing, *a way* of being slothful, *a way* of being virtuous. We have seen that Stanley's and Williamson's intellectualism fails if it hopes to provide a full reduction of knowing-how to knowing-that through practical modes of presentation, because it can only give us indirect descriptions of pieces of practical knowledge. But it might well be correct for the kinds of descriptions of practical knowledge that we get from literary works

5. A Practicalist Version of Ethopeia

It remains to be seen how the practicalist version of literary cognitivism that I propose is compatible with the view that literature can give us some general knowledge. According to my hypothesis, literary knowledge consists in practical modes of presentations or ways in Stanley's and Williamson's sense, which are both practical and expressible as propositions. But ways, on their view, are singular modes of presentation, and the pieces of practical knowledge that they consist in are not general. This was one of the main difficulties of their view: one can know various *particular* ways of ϕ-ing (say typing such or such a word on a

[17] Trump and Schwartz (1987) is supposed to tell you "the art of the deal".

[18] I am here referring to Recanati's important work on deference, elaborated in the context of his theory of reference. See e.g. his Recanati 2000.

keyboard, playing such or such a tune on an instrument, riding a bike in a given circumstance, etc.) without knowing how to ϕ, how to perform the *general* action of ϕ-ing (typing, playing a tune on an instrument, riding a bike). The practicalist version of this view suggested in the previous paragraph encounters the same difficulty. Novels, short stories, literary works of fiction and narratives of all kinds always give us descriptions that are singular, not the general case. Thus if we take up the example of stories about poisoners, Dickens' "Hunted Down", Oscar Wilde's *Pen, Pencil and Poison, A Study in Green*, both describe the figure of the poisoner Thomas Griffith Wainwright, who was a real life character. François Mauriac's, *Therèse Desqueyroux*, and many of Agatha Christie's characters also describe poisoners and their modes of operation. Do they tell us something about poisoning, its motives and practice? Certainly. Can they provide a know-how of poisoning? Certainly not, since they are descriptions of this practice in particular cases. So how could they provide the reader with some general knowledge of *how one poisons*? If we want to take seriously Joshua Landy's idea (already quotes above) that a number of fictions "rather than providing knowledge per se—whether propositional knowledge, sensory knowledge, knowledge by acquaintance, or knowledge by revelation"—give us a know-how, and that "rather than offering us a new set of beliefs, what they equip us with are skills"; if we accept his view that "rather than teaching, what they do is train", that such texts "are not informative, that is, but formative", we need to reformulate this view as a form of intellectualist practicalism. But in order to impart to us the kind of knowledge that can be formative, the knowledge imparted has to be in some sense *general* and not particular.

The knowledge of practical modes of presentation involves both singular demonstrative propositions and general propositions about laws and regularities. What kind of laws? Laws of the human mind, truths about human life, its forms, in particular its ethical forms. The modes of presentation have to involve such regularities. Here are some examples:

This is how a modern day Theresa looks (Eliot, *Middlemarch*)
Fear looks *like this* (Maupassant, *The Horla*)
Here is a way of being slothful (Goncharov*, Oblomov)*
Here is a way of being jealous (Proust, *Albertine*)
This is *the way* love looks like (Lawrence, *Women in Love)*
Here is a way of being stupid (Flaubert, *Bouvard and Pécuchet*).

In giving these examples I do not mean to imply that each novel or narrative presents only *one* way of instantiating a general and unique law. Indeed things are much more complicated. Oblomov, for example, illustrates one kind of disposition or character, and one type of vice, sloth. But, as in real life, psychological dispositions do not come one by one. They come through patterns of other dispositions, and they are manifested in many ways. They are, in Ryle's phrase, *many track*.

Practical modes of presentations allow us to recognise cases of general truths: about types of individuals and human dispositions (novels), about characters (comedy, satire, tragedy). The literary tradition of describing characters, the tradition of *ethopeia* represented by classical moralists and satirists (Theophrastus, Juvenal, or Swift) and by novelists (Fielding, Austen, Eliot, or

James), does just that.[19] The practical modes of presentations of characters in this tradition are associated to general knowledge of a psychological, sociological or historical kind, but also of a moral kind. They constitute both a form of what we might call a literary know how and of literary knowledge of laws of human psychology. Such laws, however, are never meant to be discovered like scientific laws. They are always presented, or displayed, shown within the fictions on particular cases. On this view, *ethopeia* involves:

(i) A theme (in Lamarque and Olsen's sense), a type or character, and a law (*ceteris paribus*) about human psychology.[20]
(ii) A particular example of character (say *Oblomov*).
(iii) A series of practical modes of presentation enabling us to recognize the type under that mode (these are the *exempla* of the classical moralists).

All of these constitute a set of truths, which are the objects of our propositional knowledge, although these truths are presented as singular practical modes of presentations.

6. Conclusion

A lot more should be said in order to defend fully this intellectualist version of practicalism about literary knowledge, and I have here only suggested it in outline. Several objections to this view come to mind.

First, I have said nothing about the way a reader can *interpret* the set of truths that are supposed, on this view, to be imparted by pieces of literary practical knowledge. What is shown by narration is an aspect, under a certain practical mode of presentation (which can be hugely complex: what happens in the narrative). But the communicated *know-how* is never direct. For it is deferred to the reader, who has to *interpret* it. This interpretation has to involve a lot of background knowledge, and much of this knowledge has to be the object of a form of recognition, rather than of direct cognition.[21]

Second, one might object that the form of literary cognitivism proposed here runs the risk of falling back into the trappings of propositionalism, since it insists on the fact that the literary modes of presentations are associated with the learning of general laws, which are expressible in propositions. So do not we return to a form of didacticism about literature? The same objection would insist on the idea that this intellectualist version of practicalism leaves out the form and the style inherent in literary works. And so it seems to be open to the classical objections to propositionalism presented above. The answer is that it does not. Practical literary knowledge is the acquisition of abilities to recognize characters, but these abilities are not imparted to us directly. They are represented in fiction.

A fourth objection is that the view seems to be too narrowly tied to a certain literary genre (*ethopeia* and its associated forms such as comedies, satires and novels) at the expense of others (say poetry, diaries, and contemporary nov-

[19] Carnevali 2010, Pavel 2013.
[20] On laws that one can get from literary fictions, see Elster 2010, especially his comments on Proust on self-deception.
[21] On that I agree with Gibson 2009. But I disagree with his view that the kind of understanding offered by literature is not cognitive or conceptual but only "dramatic". I agree that it is "dramatic", but deny that it is *only* such.

els). The answer is that it is not. *Ethopeia* is only the purest form of literary cognitivism. There are many others, and although they do not necessarily take this form, they may refine it. The hypothesis is that the three elements (i)-(iii) given above are always present, although not in their standard form.[22]

Fifth, one might object that I have not spelled out what I called above the aesthetic value condition of literary cognitivism: the cognitive value of literary work is essential to the aesthetic value of these works. How does one avoid what George Eliot calls "a patent deodorized and non resonant language" that "does not express life"? The answer here is to admit the shortcoming. I have not provided any explanation of the relationship between cognitive and aesthetic value. It must clearly have to do with not only the modes of presentation of the kinds of knowledge, but with the style and form in which those modes are presented. But it was not the aim of this article to articulate a full defense of practicalist literary cognitivism. My aim has been mostly to state the view and to argue that it is not inconsistent. In spite of these difficulties, I hope to have presented here the outline of an answer to the sceptic about literary cognitivism.[23]

References

Barbero, C. 2013, *Filosofia della letteratura*, Roma: Carocci.

Benda, J. 1945, *La France byzantine*, Paris: Gallimard.

Bengson, J. and Moffett, M.A. (eds.) 2012, *Knowing How: Essays on Knowledge, Mind, and Action*, Oxford: Oxford University Press.

Boswell, J. 1837, *Life of Johnson*, New York: Crocker.

Bouveresse, J. 2008, *La connaissance de l'écrivain*, Marseille: Agone.

Carnevali, B. 2010, "Mimesis littéraire et connaissance morale. La tradition de l''éthopée'", *Annales HSS,* mars-avril, 2, 291-322.

Carroll, N. 1998, "Art, Narrative and Moral Understanding", in Levinson, J. (ed.), *Aesthetics and Ethics. Essays at the Intersection,* Cambridge*:* Cambridge University Press, 126-60.

Conrad, J. 1914, "Preface", in *The Nigger of Narcissus*, New York: Doubleday.

Currie, G. 1990, *The Nature of Fiction*, Cambridge: Cambridge University Press.

Currie, G. 2010, *Narratives and Narrators: A Philosophy of Stories*, Oxford: Oxford University Press.

Currie, G. (forthcoming), "Truth and Trust in fiction", in Sullivan-Bissett, E. and Noordhof, P. (eds.), *Art & Belief*, Oxford: Oxford University Press.

[22] Pavel 2013 suggests that although this tradition is less clearly present in contemporary literature, where characters are less delineated and often disappear (or become, in Musil's phrase, *men without qualities*), it is still present nevertheless.

[23] I have read various versions of this article in conferences and seminars in Fortaleza, Madrid, Edinburgh, Paris in 2013 and 2014 and thank the organisers and audiences at these conferences, especially André Leclerc, Jesus Vega Encabo, Duncan Prichard, Claudine Tiercelin, Annick Louis, Michael Lynch, Veli Mitova, Timothy Williamson, Barbara Carnevali and Gregory Currie for their remarks. Especial thanks the participants in my 2013-14 seminar on literature and knowledge at EHESS. And I thank Massimo Dell'Utri for his insistence that I should write it up as an article for this journal.

Currie, G. and Ravenscroft, I. 2003, *Recreative Minds: Imagination in Philosophy and Psychology,* Oxford: Oxford University Press.

Dickens, C. 1999, "Preface", in *Oliver Twist*, Oxford: Oxford University Press.

Eliot, G. 1883, *Essays*, Sheppard N. (ed.), New York: Funk and Wagnall.

Elster, J. 2010, *L'irrationalité*, Paris, Seuil.

Engel, P. 2002, *Truth*, Chesham, Bucks: Acumen.

Engel, P. 2009, "Truth is One", *Philosophia Scientiae*, 13 (1), 1-12.

Engel, P. 2012, *Les lois de l'esprit*, *Julien Benda ou la raison*, Paris: Ithaque.

Engel, P. 2013a, "Trois conceptions de la connaissance littéraire: cognitive, affective, pratique", *Philosophiques*, 40, 1, 121-38.

Engel, P. 2013b (ed.), *Littérature et connaissance*, *Philosophiques*, 40, 1 (Société de philosophie du Québec).

Gaskin, R. 2013, *Language, Truth and Literature: A Defense of Literary Humanism*, Oxford: Oxford University Press.

Gibson, J. 2009, "Literature and Knowledge", in Eldridge, R. (ed.), *Oxford Handbook of Philosophy and Literature*. Oxford: Oxford University Press, 467-85.

Hazlitt, W. 1916, *The Characters of Shakespeare's Plays*, *Othello*, Quiller Couch, A. (ed.), London: Hunter and Ollier.

Horwich, P. 1990, *Truth,* Oxford: Oxford University Press.

James, H. 1884, *The Art of Fiction*, in his *Partial Portraits,* London: Macmillan, 1888, 375-405.

Johnson, S. 1765, *Preface to Shakespeare*, New York: Harvard Classics, Collier & Son, 1908.

Landy, J. 2012, *How to Do Things with Fictions,* Oxford: Oxford University Press.

Lamarque, P. 2009, *Philosophy of Literature*, Oxford: Wiley-Blackwell.

Larmarque, P. and Olsen, S. 1989, *Truth, Fiction and Literature*, Oxford: Oxford University Press.

Leavis, F.R. 1948, *The Great Tradition*, London: Chatto & Windus.

Lewis, D. 1978, "Truth in Fiction", *American Philosophical Quarterly*, 15, 37-46.

Lewis, D. 1990, "What Experience Teaches", in Lycan W., *Mind and Cognition: A Reader*, Oxford: Blackwell, 499-518.

Lombardo, P. 2012, "Comme une fiction: empathie et expérience de pensée", in Gagnebin M. and Milly J. (eds.), *Michel de M'Uzan ou le saisissement créateur*, Seyssel France: Champ Vallon.

Lynch, M.P. 2009, *Truth as One and as Many*, Oxford: Oxford University Press.

Murdoch, I. 1997, "Philosophy and Literature: a Conversation with Bryan Mage", in Murdoch I., *Existentialists and Mystics,* London: Penguin, 1-3.

Nemirow, L. 1990, "Physicalism and the Cognitive Role of Acquaintance," in Lycan W. (ed.), *Mind and Cognition: A Reader*, Oxford: Blackwell, 490-99.

Nussbaum, M. 1990, *Love's Knowledge. Essays on Philosophy and Literature*, Oxford: Oxford University Press.

Pavel, T. 2013, *The Lives of the Novel*, Princeton, NJ: Princeton University Press.

Poston, T. 2009, "Know How to Be Gettiered?", *Philosophy and Phenomenological Research*, 79, 3, 743-47.

Putnam, H. 1978, *Meaning and the Moral Sciences*, London: Routledge.

Rorty, R. 1991, *Objectivity, Relativism and Truth*, *Philosophical Papers I*, Cambridge: Cambridge University Press.

Rumfit, I. 2003, "Savoir faire", *Journal of Philosophy*, 100, 3, 158-66.

Ryle, G. 1945-1946, "Knowing How and Knowing That", *Proceedings of the Aristotelian Society*, 46, 1-16.

Ryle , G. 1949, *The Concept of Mind*, London: Hutchinson.

Stanley, J. 2011, *Knowing How*, Oxford: Oxford University Press

Stanley, J. 2011a, "Knowing (How)", *Nous*, 5, 2, 207-38.

Stanley, J. and Williamson, T. 2001, "Knowing How", *Journal of Philosophy*, 98, 8, 411-44.

Stolnitz, J. 1992, "On the Cognitive Triviality of Art", *British Journal of Aesthetics*, 32, 3, 191-200.

Trilling, L. 1950, *The Liberal Imagination: Essays on Literature and Society*, New York: Viking Press.

Trump, D. and Schwartz, T. 1987, *The Art of the Deal*, New York: Random House.

Van Delft, L. 2012, *Les moralistes, une apologie*, Paris: Folio.

Walton, K. 1993, *Mimèsis as Make Believe*, Cambridge, MA: Harvard University Press.

Wiggins, D. 2012, "Practical Knowledge, Knowing How To and Knowing That", *Mind*, 121, 481, 97-130.

Wright, C. 1993, *Truth and Objectivity,* Oxford: Oxford University Press.

Relativism, Faultlessness and Parity: Why We Should be Pluralists about Truth's Normative Function

Filippo Ferrari

Universität Bonn

Abstract

Some philosophers, like Mark Richard and Paul Boghossian, have argued against relativism that it cannot account for the possibility of faultless disagreement. However, I will contend that the objections they moved against relativism do not target its ability to account for the possibility of faultless disagreement per se. Rather, they should be taken to challenge its capacity to account for another element of our folk conception of disagreement in certain areas of discourse—what Crispin Wright has dubbed *parity*. What parity demands is to account for the possibility of coherently appreciating, within a committed perspective, that our opponent's contrary judgement is somehow on a par with our own judgement. Understood in this way, Boghossian's and Richard's objections put indeed considerable pressure on relativism—or so I will argue. I will consider John MacFarlane's attempt to resist their objections and I will show that, once their arguments are properly understood as targeting parity, the attempt is not successful. In the last section of the paper I will offer a diagnosis of what is at the heart of the relativist inability to account for parity—namely its assumption of a monistic conception of the normativity of truth.

Keywords: Truth, Relativism, Faultless Disagreement, Parity, Normative Pluralism.

1. Introduction

Anna and Marco decide to go to a new sushi restaurant downtown. They are both food lovers and they have had many past experiences of sushi together. Moreover, let us suppose that they have an impressive record of past agreements concerning the taste of sushi. On this occasion, however, Anna judges the sushi to be delicious while Marco disagrees, judging it to be not delicious. Quite surprised by their divergent judgements, they try the sushi again and yet they stick to their original judgements—Anna judging it to be delicious while Marco judging it to be not delicious. Given their backgrounds, they take this divergence in judgements at face value. In fact, they take themselves to be disagreeing about

Argumenta 2, 1 (2016): 77-94
ISSN 2465-2334

DOI 10.14275/2465-2334/20163.FER

whether a particular piece of sushi is delicious or not. However, reflecting on the subject matter of their disagreement—what we might call *disagreement about basic taste*—they also believe that nothing important hinges on it. In fact, they think that because basic taste is such a subjective matter there is no sense in which the disagreement has to be settled by determining who is right and who is wrong. In other words, they believe that in such a situation there is no sense to be made of the idea that someone has to be mistaken in judging the way she does.

I take this piece of fiction to be a philosophically informed description of a possible scenario in which Anna and Marco disagree *faultlessly*. It does not matter whether Anna and Marco qua non-philosophers would describe their situation in the way I have just done, or whether they would immediately agree on such a description. We (philosophers) know well that folk are not used to make the kind of fine-grained distinction philosophers are acquainted with. All that is required here is that the description of the exchange between Anna and Marco I have just given has a certain initial degree of intuitive grip on us—qua competent speakers of English. And, for philosophers, what matters is to come up with a coherent and satisfactory theory of basic taste that explains, or explains away, the intuitive grip that the exchange between Anna and Marco, as just described, possesses.

Various proposals have been defended in the recent debate—various forms of contextualism, hybrid-expressivism, relativism, invariantism, etc.—to deal with the phenomenon of faultless disagreement. In this paper, I will not take a stand on which theory better explains, or explains away, the phenomenon.[1] What I would like to do, instead, is to discuss whether a certain form of relativism—in fact a neutral version between what MacFarlane dubs *non-indexical contextualism* and his own *assessment-sensitive relativism* (MacFarlane 2014)—can explain the phenomenon. In so doing, I will discuss a recent exchange between four philosophers whose respective works have shaped the debate on faultless disagreement: Paul Boghossian, John MacFarlane, Mark Richard and Crispin Wright. I will argue that while MacFarlane wins the battle against the charge pressed by Richard on the adequacy of relativism to explain faultless disagreement, it is not clear whether he also wins the war of making full sense of faultless disagreement. This is because, I will contend, it is not clear whether MacFarlane can satisfactorily explain an important aspect of that phenomenon—what Wright has called *parity*. The conclusion will then be that assessment-sensitive relativism might not be able to account for the full intuition behind faultless disagreement.[2]

2. Faultless Disagreement

Max Kölbel, who initiates the recent debate on faultless disagreement, defines the phenomenon in the following way:

> A faultless disagreement (FD) is a situation where there is a thinker A, a thinker B, and a proposition (content of judgment) p, such that: (a) A believes (judges)

[1] See Ferrari 2016b, Ferrari and Wright (forthcoming).

[2] To be fair, MacFarlane does not motivate his relativism by attempting to accounting for faultless disagreement, even though he is not completely insensitive to the issue.

> that p and B believes (judges) that not-p; (b) Neither A nor B has made a mistake (is at fault) (Kölbel 2003: 53-54).

How should we understand this characterization of faultless disagreement? In particular, how should we interpret the two key notions of *disagreement* and *fault (or mistake)*? Concerning the notion of disagreement, things are not so simple. There is an ongoing discussion about what notion of disagreement we should consider in evaluating the explanatory adequacy of competing semantic theories.[3] As a result of that discussion, many have been persuaded that there is in fact a plurality of *explananda*, and thus that it is somehow misleading to talk of disagreement *as such*. However, for the purposes of this paper we do not need to engage in that debate. In fact, given the basic semantic assumptions that relativists make about the truth-aptness of judgments in the target domains, the kind of disagreement which is at issue in the characterization above is what we might call *propositional disagreement*.[4] Two thinkers—Anna and Marco—propositionally disagree just in case Anna's judgement entails the negation of Marco's judgement (alternatively, just in case Anna believes a proposition that entails the negation of what Marco believes).[5] With this clarification in hand, the phenomenon I am primarily interested in concerns the possibility of *propositional disagreements* in which neither thinker is at fault.

The notion of fault as well is open to many different interpretations. However, as for the case of the notion of disagreement, given certain background assumptions that relativists generally make, we can safely assume that the technical sense involved in the characterization of faultless disagreement above is a normative one. A thinker is committing a fault in believing a proposition *p* if and only if in so believing she is violating the relevant norms governing enquiry. What the *relevant* norms are is a matter of dispute. The three 'usual suspects' are truth, knowledge and justification.[6] In this paper I will assume, with Kölbel and MacFarlane, that the main norm governing beliefs and enquiry is truth. Thus, we have something like the following general normative constraint governing enquiry:

[3] See, for instance, MacFarlane 2014, Huvenes 2012 and Baker 2013. See also Ferrari 2016b.

[4] There might be other kinds of disagreement as when, for instance, I am entertaining a dispute with my former self in relation to a situation where there is a judgement made at t1 and later, at t2, a retraction of that very judgement. Or there might be a situation where a judgement is opposed by another subject simply by rejecting it (or suspending judgement about it). A more encompassing notion of disagreement in terms of opposition of doxastic attitudes might be required in order to account for these cases as well. However, for the purposes of this paper, the simple characterisation of disagreement in terms of the semantic incompatibility of the propositions involved in the disagreement will do. A novel kind of pluralism about disagreement is developed in Moruzzi, S. (ms).

[5] This naïve characterisation of propositional disagreement has to be refined in order to exclude cases of conflicting temporally/locationally neutral propositions as cases of genuine disagreement. See Ferrari 2016b.

[6] Advocates of the truth-norm are, among many others, Weiner 2005, Shah and Velleman 2005. For a defence of the knowledge-norm see especially Williamson 2000, Ch. 11; Hawthorne 2004, and Smithies 2012. For a defence of the justification (or reasonableness) norm, see Lackey 2007 and Kvanvig 2009.

(TR) A thinker T is correct to believe (or assert) that *p* if and only if *p* is true.[7]

With this in hand, we can rephrase our general definition of *normative fault* in the following way:

(NF) A thinker is making a mistake[8] in believing a proposition *p* if, and only if, believing *p* is deemed incorrect by (TR)—i.e., if, and only if, *p* is not true.

That said, the relativist project is to offer a conception of relative truth which is able to account for the possibility of situations in which Anna believes a proposition *p—that this sushi is delicious*—Marco believes a proposition *q—that this sushi is not delicious*—which entails *not-p*, and neither Anna nor Marco are violating (TR).

3. The Simple Deduction

There is a simple argument to show that propositional disagreement and faultlessness, in the way we have characterized these notions, are inconsistent. Such an argument is known in the literature as *The Simple Deduction*, and it goes as follows:

1. A accepts P	[Assumption]
2. B accepts not-P	[Assumption]
3. A's and B's disagreement involves no mistake	[Assumption, FD]
4. P	[Assumption]
5. B is making a mistake	[2, 4, TR, NF][9]
6. Not-P	[4, 5, 3 *RAA*]
7. A is making a mistake	[1, 6, TR, NF]
8. It is not the case that A's and B's disagreement involves no mistake	[3, 5, 7 *RAA*].[10]

Line (3) in the argument above is meant to capture the assumption that the disagreement in question is faultlessness, and the conclusion of the argument is a disproof of that assumption—namely, that it is not the case that A's and B's disagreement involves no mistake. Assuming classical logic, from line (8) we can validly infer that either A is making a mistake or B is making a mistake. Propositional disagreement precludes normative faultlessness. What can be said in reply to this argument to rescue the possibility of faultless disagreements?[11] Alethic

[7] One point which is worth mentioning is that, although I am here considering the 'if and only if' version of the truth-rule, for those that have quibbles with the 'if' direction, the rule could be restricted to the 'only if' direction without compromising the main points of the paper.

[8] I will use the notion of fault and that of mistake interchangeably.

[9] Strictly speaking step 5 follows from step 4 in virtue of these further steps:

4.1. P is true (4, T-schema)
4.2 Not-P is not true (4.1, bivalence).

The same, *mutatis mutandis*, holds for the inference from 6 to 7. Thanks to an anonymous referee for pointing this out.

[10] Wright 2006: 41.

[11] One possibility is to reject classical logic for those domains where the phenomenon of faultless disagreement strikes as plausible. This line of reply is explored by Wright 2001, 2006. But see Shapiro and Taschek 1996 and Shapiro 2012 for some challenges.

relativism seems to offer a very neat and simple solution—by relativizing truth they also relativize the judgement-truth norm and, consequently, the notion of being at fault. The general thought is that a judgement that p is in good standing if and only if p is true relative to the subject's perspective.[12] Thus, a subject is at fault just in case she judges something false relative to her perspective.

4. Alethic Relativism

In recent philosophy of language two main versions of alethic relativism have crystallized: a moderate version, championed by Max Kölbel, and a more radical one recently defended by John MacFarlane. The view has been applied to various domains of discourse, including the domains of taste, epistemic modals, knowledge attributions, and the moral domain. For simplicity and easiness of exposition, in this paper I will focus exclusively on the application of relative truth to judgments of taste (e.g., judgments like "this sushi is delicious").[13]

Both varieties have much in common, in particular they both take judgments in the taste domain to express truth-apt contents whose truth value is relative to the taste sensibility of an agent (either that of the speaker or that of the assessor). Despite these important similarities, there is a fundamental semantic difference between the two views, which is what makes MacFarlane's relativism *more radical* than Kölbel's. Assuming a broadly Kaplanian approach to semantics, such a difference consists in the fact that besides introducing non-standard parameters into the circumstances of evaluation—e.g., those tracking an agent's taste sensibility—MacFarlane introduces a non-standard context, which he calls the *context of assessment*. The specific semantic function of such a context is that of providing the default value for the various parameters in the circumstances. Whereas in Kaplan's original framework, as well as in Kölbel's more conservative extension of it, the context of use provides the default information for the evaluation of the truth of a sentence-in-context, in MacFarlane's framework the context of assessment fulfils that function. Although according to MacFarlane this difference is crucial to account for what he calls the *retraction phenomenon*, it won't matter for our purposes. This is because, when it comes to the normative significance of disagreement and the norms governing judging, the two frameworks give similar predictions (MacFarlane 2014: 102-106). The reason is that when we focus on judging, or the making of assertions, context of use and context of assessment coincide.[14] An agent, in judging that p, is also assessing p's truth relative to the context in which the judgement is performed. If we consider only the normative consequences of judging, we cannot tell the difference between the two theories. Thus, we should expect that the two make analogous predictions concerning the normative aspects of disagreement. In this respect, we can treat the two theories as on a par for the purposes of this paper. In what follows I will use the term 'relativism' in such a way as to cover both non-

[12] I am using the general notion of a perspective to be neutral as to whether the relevant context to which relativize truth is that of utterance or that of assessment.

[13] I am here focusing only on particular judgements of taste—i.e. judgements of the form "this (particular) piece of sushi is delicious". General judgements of taste—i.e. "sushi is delicious (in general)" introduce an additional layer of complexity which can be avoided for present purposes.

[14] MacFarlane focuses on assertions, but for present purposes it will not matter much.

indexical contextualist variants as well as assessment-sensitive ones—call this *minimal relativism*.

What this minimal relativism amounts to—in the case of basic taste—is the idea that the truth of judgements in that domain is relative to the taste perspective[15] of a subject—either the judge or the assessor. Thus, in this sense, truth is perspectival—it does not make sense to ask for whether a certain taste judgement is true independently of any given taste-perspective.

For reasons that will become clear in subsequent sections, it is important to note that relativists also allow for a non-relative, fully disquotational monadic truth predicate that operates within a given perspective—call this *truth simpliciter*. Once a subject is within a taste-perspective, she can make use of the truth simpliciter to make non-relative truth-ascriptions. This fact, as we will see, is going to be crucial in discussing the parity element of the faultless disagreement phenomenon.

5. Relativism and the Simple Deduction

How does relative truth help with respect to faultless disagreement and the Simple Deduction? By relativizing truth relativists also relativize (TR):[16]

(Rel-TR) A thinker A is correct to judge that *p* if and only if *p* is true relative to A's perspective.

Since *that this sushi is delicious—p—is* true relative to Anna's perspective, and *that this sushi is not delicious—q—*is true relative to Marco's perspective, we have that Anna is correct in judging that *p* and Marco is correct in judging that *q*. In this way, both Anna and Marco are in compliance with (Rel-TR). Yet, they disagree since in the relativist framework we still have that Marco's judgement contradicts Anna's judgement. The Simple Deduction is thus effectively blocked at the step from 4 to 5. Hence—putting worries concerning what it means to follow a relativised judgement-norm aside—[17]at least with respect to the domain of judgments of taste, the logical possibility of faultless disagreement is accounted for.

[15] I am using the general notion of a *perspective* in order to be neutral with respect to the issue whether the relevant context to which relativize truth is that of utterance or that of assessment. Cf. Boghossian 2011: 65, Cappelen and Hawthorne 2009, Ch. 4.

[16] One might think that it is not obvious that the relativist should move straightforwardly from TR to Rel-TR. It is at least conceivable that there be a relativist who, for some relativist judgement p, maintains TR and consequently judges anyone who believes not-p to be at normative fault—even if according to the standards of the interlocutor, not-p is true. There is also a view potentially open to the relativist according to which "person A is normatively at fault in believing not-p" is *itself* a relativist judgement, one that is true in the mouth of someone whose standards endorse p, and false in the mouth of others. To see that this differs from Rel-TR, note that according to Rel-TR it is an absolute matter whether someone is normatively at fault in believing that p, while on this account it is a relative matter. However, neither options, although viable relativistic alternatives, help with the issue of parity. Many thanks to Dan Waxman for a discussion on this point.

[17] On this point, see Marques 2014. Moreover, on the issue concerning whether it is rational to believe in a relativized judgement norm see Moruzzi 2009.

6. Faultlessness & Parity

Even conceding to relativists that they have the tools to resist the Simple Deduction, one might argue that there is still an important aspect of the general intuition concerning disagreement about matters of taste that is left unexplained. What we are looking for is not only a demonstration of the logical consistency of propositional disagreement and normative faultlessness, but also an explanation of how such a fact can be coherently appreciated and expressed within a committed perspective taking part to the disagreement—i.e. consistently endorsed together with a thinker's own opinion on the subject matter of the disagreement. Wright calls this extra ingredient *parity*, and he characterises it as follows: "In effect, it is the requirement that faultlessness be appreciable, and endorsable, from the point of view not just of neutrals but of the committed parties in a dispute" (Wright 2012: 439). According to Wright, this feature of taste disagreement is "meant to be implicated by faultlessness—conveyed in the acknowledgment that *your opinion is just as good as mine*" (*Ibid.*, emphasis in the original).

There is an important difference between parity and faultlessness. Such a difference resides primarily on the different point of view that is involved. Whereas the evaluation of the faultlessness of a disagreement is made from within a neutral point of view—that of a referee external to the disagreement who does not take a stand on the topic of the disagreement—the evaluation of parity is carried out from within the committed perspectives of the judgers involved in the disagreement. For this reason, there is a clear sense in which accounting for parity requires more theoretical resources than accounting for the logical possibility of faultlessness. In fact, what it asks for is to make space for the permissibility within a committed perspective of a judgment to the effect that the disagreement is one that does not necessarily involve fault and, consequently, that the opponent is under no rational requirement to change her mind in the light of the disagreement.[18]

But why should a relativist care about this extra feature? Should not we be happy with an effective account for the logical possibility of faultless disagreement? Should not we just refuse to acknowledge this extra feature as part of the explanandum? This is an open possibility, but an inconvenient one. The reason why we should care about parity is that it seems to be an important part of our pre-theoretical conception of disputes about matters of taste. One might even argue that from a folk-theoretical point of view the parity feature is explanatory prior to the faultlessness feature. The idea is that what drives the more abstract thought concerning the faultlessness of certain disagreements about taste is the appreciation in many cases of actual confrontations about matters of taste that our opponent's judgement is, in some important respect, no less legitimate than mine. We want a theory of the normative significance of disagreement in the

[18] Strictly speaking, "Parity" involves two conditions: 1) that each of the disputants acknowledges the other not being at fault; 2) that they still maintain that there is a real disagreement. Another option could be to hold on a weaker notion of parity according to which it involves just the satisfaction of the first condition. After all, a way in which the disputants can acknowledge that none of theme is at fault is to recognise that their opinions have non-contradictory contents. However, this contextualist-like move seems to undermine much of the motivation for a relativist semantics (either use- or assessment-sensitive), so I do not take it to be a viable option for a relativist. Thanks to an anonymous referee for suggesting this point.

domain of taste that gives us—ordinary speakers engaging in everyday disputes about taste—the tools for accounting for this important aspect of the folk conception. Thus, if relativists want to stay in the game they must provide us with an explanation not only of the logical possibility of faultless disagreement but also of the parity feature of disagreements about taste.[19]

7. Boghossian and Richard on Relativism and Faultlessness

There are reasons to suspect that for relativists accounting for parity might not be as straightforward as accounting for the possibility of faultless disagreement. In fact, merely relativizing the truth norm does not seem to help with respect to parity. In what follows I will review two similar arguments, one from Mark Richard and the other from Paul Boghossian, which show why relativism, as it stands, fails to account for parity. I will then consider MacFarlane's attempt to address these arguments for then arguing that such an attempt might not be effective in rescuing relativism from the parity challenge.

Richard and Boghossian have put forward arguments intended to cast doubt on the relativists' ability of accounting for the possibility of faultless disagreement. In fact, because neither Richard nor Boghossian was distinguishing between parity and faultlessness, they both take their respective arguments to show that relativizing truth, and consequently the truth-norm, does not suffice for a full explanation of the possibility of faultlessness. With the distinction between faultlessness and parity in hand, we can appreciate that taken as arguments against the possibility of faultless disagreements, they both fail. However, they can be effectively used to show that relativists are in trouble in giving an effective account of parity. Thus, in this section I will outline these arguments and I will use them to show that relativists cannot account for parity.

Richard, in introducing his own view concerning matters of taste, presses the following line of criticism against alethic relativism:

> Suppose I think that Beaufort is a better cheese than Tome, and you think the reverse. Suppose (for reductio) that each of our thoughts is valid – mine is true from my perspective, yours is from yours. Then not only can I (validly) say that Beaufort is better than Tome, I can (validly) say that it is true that Beaufort is better than Tome. And of course, if you think that Tome is better than Beaufort and not vice versa I can also (validly) say that you think that it is not the case that Beaufort is better than Tome. So, I can (validly) infer that it is true that Beaufort is better than Tome though you think that Beaufort isn't better than Tome. From which it surely follows that you are mistaken – after all, if you have a false belief, you are mistaken about something. This line of reasoning is should

[19] Another reason for this conclusion is that if a theory allows for faultlessness but not for parity, then whoever comes to believe in the parity condition is in error according to the theory. Thus, such a theory prescribes that a thinker must either have no view about parity or she must believe in its falsity. But believing in the falsity of the parity condition is a way to say that the relativistic doctrine cannot be taken as a real commitment—i.e. that judging from a perspective does not allow for conceding any ground to the opponent's view, contrary to what the doctrine predicts from an abstract point of view. In both cases, the broader conclusion is thus that a failure to accounting for parity has the consequence that the relative doctrine cannot be coherently endorsed by anyone having a committed perspective on the subject matter of the dispute. Many thanks to an anonymous referee for suggesting this point.

> no matter what the object of dispute. So, it is just wrong to think that if my view is valid - true relative to my perspective – and your contradictory view is valid - true, that is, relative to yours – then our disagreement is 'faultless' (Richard 2008: 132).[20]

The upshot of Richard's argument is that, even if we endorse a relativistic conception of truth, within a committed perspective a thinker is committed to evaluate a contrary opinion as false. And for that reason, she is committed to evaluate anybody holding such an opinion as being at fault in judging in a way she ought not to. From this, Richard concludes that the disagreement is not faultless.

Boghossian challenges alethic relativism in a very similar vein, offering an argument which he calls "The Argument from (Perspectival) Immersion" (Boghossian 2011: 62). The argument—as I intend it—goes roughly as follows:

1. The content of a taste proposition p is relatively true [Def. of Relativism]
2. p is true within D's perspective and D judges that p [Assumption]
3. If D judges that p and p is true within D's perspective, then D is correct in judging that it is true that p [Truth Simpl., TR]
4. It is correct for D to judge that it is true that p [2, 3, MPP]
5. If, within D's perspective, it is true that p then, within D's perspective, it is false that not-p [Logic of Truth Simpl.]
6. Within D's perspective it is false that not-p [2, 5, MPP]
7. If within D's perspective it is false that not-p, then D is correct in judging that it is false that not-p [Truth Simpliciter, TR]
8. D is correct in judging that it is false that not-p [6, 7, MPP]
9. If D is correct in judging that it is false that not-p, then, on pain of incoherence, D is correct in judging that anyone judging not-p (e.g., N) is making a mistake [TR, NF]

Therefore,

10. D is correct in judging that N is making a mistake [8, 9, MPP]
11. D is correct in judging that her disagreement with N is not faultless [10, FD]
12. The disagreement between D and N is not faultless [1, 11, FD]

Steps from 1 to 8 seem hardly objectionable. Step 9 clearly requires some additional consideration since it is not straightforward at all why we should endorse it. I will return to it shortly.

The upshot of Boghossian's line of reasoning is that any speaker within a committed perspective can correctly evaluate anybody holding a contrary opinion as making a mistake and thus can correctly evaluate the disagreement as not faultless.

[20] Wright makes a similar point when he writes: "It is pretty immediate that Assessment-relativism is useless for the purpose of securing Parity. By its rules, I am constrained to assess your opinion in the light of my standards, rather than yours. So, of course, I will assess it as false. Since I assess my own as true, I can then, surely, hardly regard your opinion as just as good as mine, and Parity is surrendered from my point of view, the point of view of a participant in the dispute" (Wright 2012: 440). See also Boghossian 2011: 61-62.

Boghossian, like Richard, takes the major conclusion of the argument to be that the disagreement itself is not faultless. However, both the conclusion expressed by the last sentence of Richard's argument, and the step from 11 to 12 in Boghossian's argument seem objectionable. My diagnosis of why this is so has to do with the fact that they fail to distinguish between parity and faultlessness. Granted that within a committed perspective any thinker has to evaluate a contrary judgement as false and granted that—pending an argument for line 9—anybody holding such an opinion is committing a mistake, we cannot conclude, without further argument, that the disagreement *itself* is one involving some fault on the part of either disputants. What we are allowed to conclude is that both disputants are licensed, from within their respective committed perspective, to evaluate their respective opponent as making a mistake in judging the way she does. But, even if we concede the argument until line 10, a relativist has reasons to resist the step from 11 to 12. In fact, the relativist at this point would insist that their doctrine gives us the resources to say that the disagreement between Anna and Marco is faultless. With the post-semantic notion of relative truth in hand the relativist can say that both Anna and Marco judge truly and thus faultlessly relative to their respective perspectives.

In this respect, even conceding that the argument from 1 to 10 is sound, the relativist could resist the conclusion 12 by pointing out that one thing is to provide a coherent explanation of why the disagreement itself is one that does not need to involve any fault—which relativist can offer—and another, far more demanding thing, is to provide an account of how each committed party can consistently assess her opponent's contrary opinion as faultless alongside with her own view on the subject matter at issue. And this is as it should be if the distinction between the parity and the faultlessness features is a significant one.

The point I am making in this section is that once the distinction between parity and faultlessness is acknowledged it seems clear that both Richard's and Boghossian's should not be taken as arguments against relativists' ability to account for the possibility of faultless disagreement. However, they might as well work as arguments showing that relativists are ultimately unable to account for parity. Whether a position's inability to account for parity translates *eo ipso* into an inability to account for faultlessness is something that requires argument.

With this in hand, in the next sections I will scrutinize both the soundness and dialectical effectiveness of these two arguments taken as targeting parity and not faultlessness. In doing so I will discuss a recent reply by John MacFarlane to Richard's line of reasoning which, if correct, would cast doubt also to Boghossian's argument. I will argue that MacFarlane's reply is not effective against the parity objection—and this will require some discussion of step 5 in the argument from perspectival immersion.

8. MacFarlane's Reply

Quite predictably, MacFarlane attacks line 9 of my version of Boghossian's argument—i.e. it attacks the step from an attribution of falsity to a contrary judgment from within a perspective to an attribution of fault to the subject endorsing that opinion. According to MacFarlane this step is problematic because it presupposes a non-relative normative bridge principle between truth and norms governing judgement in the targeted domain. But once we take on board a nor-

mative principle broadly on the lines of (Rel-TR) according to MacFarlane, the step from 8 to 9 remains unsupported.

Moreover, in his reply to Richard's argument (MacFarlane 2012: 453), and later in his *Assessment Sensitivity* (MacFarlane 2014: 133-35), MacFarlane distinguishes between different senses of the expression "being at fault". Two senses are particularly relevant for our purposes. Under one understanding of "being at fault"—let's call it fault#1—a thinker S is at fault#1 in judging that *p* just in case in doing so she violates the constitutive norms governing judgement. Under the second understanding of "being at fault"—fault#2—S is at fault#2 in judging that *p* just in case it is not true that *p*—where the relevant notion of truth here is the intra-perspectival notion of truth simpliciter. With this distinction in hand, MacFarlane goes on to claim that although committed parties to a dispute should regard each other as being at fault#2 they need not regard each other as being at fault#1. In that respect, according to MacFarlane no sense can be made of the possibility of disagreement without fault#2, even though he thinks that within his assessment-sensitive framework sense can be made of the possibility of disagreement without fault#1. After all, it seems that appealing to the relative truth norm provides us with the tool to account for parity as well.

Unfortunately, things are not that simple. There are two sets of issues I would like to discuss: the first is whether, in fact, relativism can give us a satisfactory account of parity by appealing to the relative truth norm; the second concerns the normative relation between the two senses of 'being at fault' that MacFarlane distinguishes.

On the first point, I contend that it is not clear that appealing to the relative truth norm offers us an effective tool to account for parity. To see the point, let us reflect on Marco's contrary judgement from within Anna's committed perspective. The relative truth norm allows Anna to claim that Marco's judgement is not in violation of the norm because it is in accordance with his standard of taste. In other words, Anna can claim that relative to his own standard of taste, Marco is judging correctly. So far so good. But is this enough to assuage the parity intuition? I doubt it. The question at this point is how Anna should assess Marco's standard of taste. What should she say about a standard of taste that permits the endorsement of a belief whose content she is committed to assess as false, from within her perspective? There seem to be only two sensible options. The first option is that she takes Marco's standard of taste to be inferior to her own. Although this option copes well with the intuitive idea that she has a commitment to her own standard of taste—and thus a commitment to prefer her standard over those that permits contrary judgements—it seems to preclude a full account of parity. Anna would find herself to endorse the following predicament concerning Marco's situation: "Your judgement is correct relative to your standard but your standard is inferior to mine". How can a judgement that is issued from an inferior standard—albeit correctly so—being on a par with a judgement that is correctly issued by a superior standard? It seems that the sense of parity that can be recovered from the relative truth rule in this scenario is rather flimsy and does not offer a satisfactory account of our pre-theoretical conception of parity. The second option available to Anna is to take Marco's standard of taste to be on a par with her own. In this respect relativism could give us a more substantive account of parity. Not only Anna is in a position to assess Marco's contrary judgement as correct relative to his own standard—she is also in a position to claim that because his standard is as good as her own the two

judgements are really on a par. Parity would be accounted for. However, I wonder whether it is fully coherent for Anna to evaluate as equally good a standard of taste which permits the endorsement of a belief whose content contradicts that of Anna's belief, alongside with a full commitment to her own standard—and thus to her own judgement. In other words, attributing full credit and equal good-standing to a standard that permits the formation of a contrary judgement seems in tension with the idea that we have a full commitment to our own standard. This is no conclusive objection against relativism, but it is a call for a more detailed explanation of what a standard of taste is and, in particular, what are the normative consequences of endorsing a certain standard in terms of the assessment of contrary standards. Until we have a more detailed story about these issues, it is not clear that relativists can give us a satisfactory account of parity.

I now turn to a discussion of the second point—i.e. a discussion of what is exactly the relation between the two senses of 'being at fault' distinguished by MacFarlane. To be honest, I am not entirely sure what to make of this distinction. In particular, it is not clear to me what sense can be made of a notion of fault—i.e. fault#2—linked to a notion—intra-perspectival truth—which plays no role in what MacFarlane takes to be the constitutive norms governing judgement. Given that the notion of fault is a notion intimately connected with normative evaluation, it is hard to understand what is the intended sense of fault#2. In other words, fault#2 can indeed be understood as a notion of fault only if intra-perspectival truth and falsity are taken to be somehow linked to the constitutive norms governing judgements. But this manoeuvre would reopen the question whether relativists can address the parity objection. In particular, something should be said about what exactly is the relation between these two notions of fault in connection with their normative significance. Given that they are both tied to the normative assessment of judgements one might wonder which of these notions have normative priority. I will return to this point in the next section.

The easiest option here for MacFarlane would be to deny that the intra-perspectival notions of truth and falsity carry any normative punch. In particular, he would have to deny that an intra-perspectival attribution of falsity to a contrary judgement licenses any attribution of normative fault. This would amount to claim that fault#2 is not really an interesting notion of fault. The trouble here is that, as we will see in the next section, there seems to be independent reason to maintain that intra-perspectival truth and falsity—even when construed in a purely deflationary fashion—function properly as norms of judgement. If that is correct, any attribution of falsity within a committed perspective engenders an attribution of fault. Hence the parity objection would still be effective.

9. Parity, the Normativity of Truth and Alethic Relativism

That truth always plays a normative function on judgement is established by an argument originally given by Crispin Wright against the deflationary conception of truth. Deflationists claim that since truth has no nature it cannot be a normative notion. Wright shows that by using the very same principles that deflation-

ists accept—foremost the equivalence schema[21] and the thesis that truth is primarily a device for expressing endorsement of a proposition or a collection of propositions—together with some uncontroversial assumptions about the logic of negation and the biconditional, we can prove that truth is a norm of judgement.

Wright's argument comes in two stages. The first stage establishes that truth and justification coincide in terms of positive normative force. A reason to regard a proposition as justified is a reason to endorse it as belief, and conversely. Moreover, a reason to endorse a proposition as belief is, by the equivalence schema, a reason to regard the proposition as true, and conversely. Thus, a reason to regard a proposition as justified is a reason to regard it as true, and conversely (Wright 1992: 18). This establishes that both truth and justification are norms of judgement.

The second stage of the argument is purported to show that truth and justification are different norms of judgement. This has to do with the fact that truth and justification potentially diverge in extension. Intuitively, a proposition can be true without being justified and, conversely, a proposition can be justified without being true. Formally, this can be shown by first noticing that truth commute with negation within the scope of the biconditional in the equivalence schema. In other words, it can be shown that truth satisfies the following negation equivalence:

(NE) <p> is not true if and only if <not p> is true.[22]

However, the corresponding principle with justification does not hold:

(NE-J) <p> is not justified if and only if <not-p> is justified.

This failure is due to the possibility of neutral states of information—a subject S is in an epistemically neutral situation with respect to her informational state i and to a proposition <p> just in case i provides S with neither a justification for <p> nor a justification for <not-p>. Relative to an epistemically neutral state of information, both <p> and <not-p> fall outside the extension of 'justified'. Moreover, if independently of any state of information either <p> or <not-p> fall nonetheless in the extension of 'true', we can infer that truth and justification potentially diverge in extension. Because of this potential divergence in extension, truth and justification cannot be the same norm. Thus, truth has to be a *sui generis* norm of judgement—i.e. a norm of judgement independent of justification. Intuitively, there is one kind of bad-standing in judging that p when p is not justified relative to the subject's informational state—regardless of whether p is true or not. And there is a *different* kind of bad-standing in judging that p when p is false—regardless of whether the subject has justification for judging that p.[23]

[21] In its propositional form, the schema is as follows: (ES) <p> is true if and only if p—where '<>' is a device for referring to the proposition expressed by the sentence encapsulated, and 'p' is a schematic letter for a sentence. Like Horwich (1998), we understand the biconditional in (ES) as a material biconditional.

[22] *Proof*: (1) <p> is true if, and only if, p [equivalence schema]; (2) <p> is not true if, and only if, not p [from 1 by contraposition]; (3) <not p> is true if, and only if, not p [from 1 by substituting <p> with <not-p>]; (4) <p> is not true if, and only if, <not p> is true [from 2 and 3 by transitivity].

[23] For a more detailed discussion of this point see Ferrari 2016c and, especially, Ferrari and Moruzzi (ms1) and (ms2).

The upshot of this for our discussion is that as soon as we endorse the basic commitments about truth that deflationists are happy to endorse, we cannot deny that truth plays a normative role with respect to judgements. Since there is no reason to deny that the intra-perspectival, fully transparent, notion of truth that MacFarlane introduces in what he calls the semantic proper obeys to the basic commitments characterising the deflationary conception of truth, there's no reason to deny that such a notion exert a sui generis normative constraint on judgement. This means that there must be, after all, a normative sense of being at fault (being at fault#2) that goes hand in hand with an attribution of intra-perspectival falsity. The question of whether relativism can account for parity is thus still open.

If this line of thought is correct, it puts considerable pressure on relativists to say something more about what is the relationship in terms of normative function between the truth-simpliciter norm and the relative-truth norm. In particular, it seems that the above arguments show that MacFarlane is committed to claim that in terms of normative significance the relative-truth norm always trumps the truth-simpliciter norm. Although I believe that this is, in principle, a viable option, some argument is required on the relativist side in order to show that this line is ultimately stable. That said, in what follows I would like to briefly explore an alternative route. I will assume for the time being that both the relative-truth norm and the truth-simpliciter norm function normatively over judgements, and they do so independently of each other. One could motivate the need of a relative-truth norm over and above the truth-simpliciter norm on pragmatic grounds—as MacFarlane seems to suggest (MacFarlane 2014, Ch. 5). But then the pressing question would be: how should we interpret the normative function of the truth-simpliciter norm in such a way to allow for a decent notion of parity? An answer to this question will be the topic of the next section.

10. What Kind of Fault?

To briefly recap: what the above discussion shows us is that if we take judgements of taste to express truth-apt contents which are, modulo standard indexicality, semantically invariant across contexts of use/assessment—call this *minimalist taste semantics*—then we have no option but to say that a commitment to assess a contrary judgement as false is *ipso facto* a commitment to attribute (some kind of) fault to any subject endorsing it. The nature of this error-attribution is epistemic—or, more properly, alethic given the normative contrast just noticed between truth and justification.

The crucial question at this point is: how serious is this attribution of fault? Generally, the normative function that truth exerts on judgements is taken to be quite substantive—in fact, to express a deontic requirement. In particular, some of the philosophers who take truth to be the norm of judgement would endorse the claim that in judging falsely a thinker is doing something (alethically) impermissible. MacFarlane is no exception to this trend.[24] In fact, he cashes the truth-rule out in terms of permissibility/impermissibility to the effect that if S judges that p and p is untrue then S is doing something impermissible (MacFarlane 2014: 103). However, understanding truth's normative function rigidly in deontic terms seems to preclude the possibility of rescuing an interesting notion

[24] See also Gibbard 2005 and Wedgwood 2007.

of parity in some domains—e.g. the domain of taste. This is because a commitment to assess a contrary judgement as false would engender—regardless of which domain that judgement belongs to—a commitment to attribute substantive fault—i.e., an impermissibility-entailing type of fault—to anybody holding that judgement. But this seems utterly too strong a prediction in the case of basic taste.[25]

Thus, in order to account for an interesting notion of parity in the taste domain we need to weaken the sense of fault which can be legitimately attributed to a contrary view on matters of taste. In other words, what we need is a way of defending the thesis that no *substantive* fault need to be attributable to a subject holding a view we are deemed to assess false. If this can be done, a decent sense of parity can be rescued. More precisely, one could argue that if sense can be made within a minimalist taste semantics of the following combination of thoughts—your judging that not-p is incorrect, because false, but there is no sense in which your judgement is any worse than mine or a judgement you ought not to have—then we could claim that there is a good sense of parity that can be preserved, despite the arguments discussed above. In what follows I will briefly outline how a proposal along these lines can be developed.[26]

The first step is to notice that when we say that truth exerts a normative constraint on judgement we could mean one or more of the following things:

CRITERIAL	It is correct (fitting) to judge that p (if and) only if p is true.
AXIOLOGICAL	It is valuable (good) to judge that p (if and) only if p is true.
DEONTIC	One ought to judge that p (if and) only if p is true.

These are three distinct dimensions of the normative constraint that truth can exert on judgment.[27] With this in hand, I call a normative alethic principle any principle expressing the normative constraint that truth exerts on judgment in terms of one or more of the three aforementioned dimensions—i.e. criterial, axiological, and deontic. It is important to appreciate the fact that the notion of a normative alethic principle so defined allows for some flexibility. In this sense, the account sketched here can be properly seen as a form of *normative pluralism* concerning truth's normative function.

The second step is to maintain that parallel to the threefold distinction in truth's normative function we have a plurality of ways in which someone holding a view that is judged untrue might be said to be at fault. Thus, we have the following three categories of attribution of fault:

DEONTIC FAULT	In judging not-p the subject is judging in a way she ought not to.
AXIOLOGICAL FAULT	In judging not-p the subject is doing something disvaluable.
CRITERIAL FAULT	In judging not-p the subject is judging incorrectly.

Although I will not argue for this here (See Ferrari (ms)), I assume that these three categories of attribution of fault are independent of each other—in particular, that criterial fault does not entail either deontic or axiological fault. With

[25] For similar consideration concerning the normative significance of retraction in various domains of discourse, see Ferrari and Zeman 2014.
[26] For a more detailed account and defence of how this should be done, see Ferrari 2016a and especially Ferrari, F. (ms).
[27] For a defence of this point see Ferrari 2016a and Ferrari (ms).

this in hand, we have open the possibility of there being a domain of discourse—e.g. basic taste—in which truth's normative function is limited to the criterial aspect and thus in which the only legitimate attribution of fault is the criterial one. This would mean that although Anna (Marco) is committed to assess Marco's (Anna's) contrary judgement as false and incorrect she (he) can nevertheless coherently claim that Marco's (Anna's) judgement is in no sense worse than her (his) own, nor a judgement he (she) ought not to have. In this way, sense can be made of the thought that when it comes to matter of taste no opinion is either impermissible or any worse than any other—provided, of course, that such an opinion sincerely reflects the author's gustatory sensibilities. An interesting sense of parity with respect to matters of taste can thus be rescued, consistently with our minimalist taste semantics.

11. Conclusions

Where does this leave us with respect to alethic relativism? I have argued that by introducing a post-semantic relative notion of truth, relativists are in a position to account for the possibility that a certain kind of disagreement about matters of taste is faultless. From a standpoint which is neutral with respect to the subject matter at issue in the disagreement—let us say, the standpoint of an uncommitted relativist—having relative truth in the theoretical toolkit allows us to say that neither party has to be at fault.

However, I have also argued with Boghossian and Richard that relativism might not deliver us everything we want. There is an element at the core of the faultlessness intuition about basic taste—what Wright has called *parity*—which relativists might have troubles in accounting for. I have claimed that we should keep questions concerning the possibility of faultless disagreement distinguished from questions concerning the possibility of parity. With this distinction in hand I have argued that both Boghossian's and Richard's arguments might be effective as arguments against relativists' inability of accounting for parity—even though they fail as arguments against faultlessness.

I have then relied on an argument given by Wright in *Truth and Objectivity* to draw a general lesson from the exchange between MacFarlane, Boghossian and Richard—namely that if we take judgements of taste to express truth-apt contents which are, modulo standard indexicality, invariant across contexts of use/assessment, then we have no option but to say that a commitment to assess a contrary judgement as false is *ipso facto* a commitment to attribute (some kind of) error to any subject endorsing it. Thus, a full notion of parity seems precluded by our minimal semantic assumption.

However, I have argued that we should not despair—that an interesting notion of parity can be rescued once we introduce some distinctions concerning truth's normative function on judgement and thus concerning the kind of fault legitimately attributable, from within a committed point of view, to someone holding a contrary judgement. I have offered a brief outline to show how this can be done and I have concluded that such a proposal can indeed give us an interesting notion of parity—which is, I think, all we can hope for. Moreover, and crucially, this proposal can be implemented within an assessment sensitive framework and would offer relativists an effective tool to assuage the parity objection alongside with a tool, provided by the post-semantic relative notion of truth, to account for faultlessness.

That said, a crucial question remains whether relativism is still needed after all, or whether we can do a good enough job with a minimalist semantics, in line with that elaborated by Wright in *Truth and Objectivity*, coupled with the normative pluralism framework offered here.[28] However, I must leave a discussion of this point for another occasion.[29]

References

Baker, C. 2013, "The Role of Disagreement in Semantic Theory", *Australasian Journal of Philosophy*, 1, 1-18.

Boghossian, P. 2011, "Three Kinds of Relativism", in Hales, S. (ed.), *A Companion to Relativism*, Oxford: Blackwell, 53-69.

Cappelen, H. and Hawthorne, J. 2009, *Relativism and Monadic Truth*, Oxford: Oxford University Press.

Coliva, A. (ed.) 2012, *Mind, Meaning and Knowledge: Themes from the Philosophy of Crispin Wright*, Oxford: Oxford University Press.

Ferrari, F. 2016a, "Disagreement about Taste and Alethic Suberogation", *Philosophical Quarterly*, 66, 264, 516-35.

Ferrari, F. 2016b, "Assessment Sensitivity", *Analysis* (doi: 10.1093/analys/anw021).

Ferrari, F. 2016c, "The Value of Minimalist Truth", *Synthese* (doi:10.1007/s11229-016-1207-9).

Ferrari, F. (ms), "Normative (Alethic) Pluralism", unpublished manuscript.

Ferrari, F. and Moruzzi, S. (ms1), "Deflationary Pluralism", unpublished manuscript.

Ferrari, F. and Moruzzi, S. (ms2), "Deflationism, Inflationism and Alethic Pluralism", unpublished manuscript.

Ferrari, F. and Wright, C. (forthcoming), "Talking with Vultures", *Mind* (doi: 10.1093/mind/fzw066)

Ferrari, F. and Zeman, D. 2014, "Radical Relativism and Retraction", in Bacchini F., Caputo, S. and Dell'Utri, M. (eds.), *New Frontiers in Truth*, Newcastle upon Tyne: Cambridge Scholar Publishing, 80-102.

Gibbard, A. 2005, "Truth and Correct Belief", *Philosophical Issues*, 15, 338-50.

Hawthorne, J. 2004, *Knowledge and Lotteries*, Oxford: Oxford University Press.

[28] See Ferrari and Wright (forthcoming) and Ferrari (ms).

[29] Acknowledgements: I would like to thank Sebastiano Moruzzi and Erik Stei for their extensive written comments on the penultimate version of this paper. A first version of it was written during my tenure of a postdoctoral fellowship in the Leverhulme Trust funded project, "Relativism and Rational Tolerance"—held at the former Northern Institute of Philosophy in Aberdeen—I would like to thank everybody at NIP for the many occasions in which they have helped me and supported my research. In particular, I would like to thank Patrick Greenough, Nikolaj Pedersen, Eva Picardi and Crispin Wright for insightful discussions on these topics over the past few years. The final version of this paper has been written during my tenure of a postdoctoral fellowship at the University of Bonn, within the project "Disagreement in Philosophy", sponsored by the German Research Foundation (DFG—BR 1978/3–1). I gratefully acknowledge the support of these funding bodies. Thanks also to two anonymous referees from *Argumenta* for their precious comments.

Horwich, P. 1998, *Truth* (2nd Edition), Oxford: Oxford University Press.

Huvenes, T. 2012, "Varieties of Disagreement and Predicates of Taste", *Australasian Journal of Philosophy*, 90, 1, 167-81.

Kölbel, M. 2003, "Faultless Disagreement", *Proceedings of the Aristotelian Society*, 105, 53-73.

Kvanvig, J. L. 2009, "Assertion, Knowledge, and Lotteries", in Pritchard, D. & Greenough, P. (eds.), *Williamson on knowledge*, Oxford: Oxford University Press, 140-60.

Lackey, J. 2007, "Norms of Assertion", *Noûs*, 41, 4, 594-626.

MacFarlane, J. 2012, "Richard on Truth and Commitment", *Philosophical Studies*, 160, 3, 445-53.

MacFarlane, J. 2014, *Assessment Sensitivity: Relative Truth and its Applications*, Oxford: Oxford University Press.

Marques, T. 2014, "Relative Correctness", *Philosophical Studies*, 167, 361-73.

Moruzzi, S. 2008, "Assertion, Belief and Disagreement: A Problem for Truth-Relativism", in García-Carpintero, M. and M. Kölbel (eds.), *Relative Truth*, Oxford: Oxford University Press, 207-24.

Moruzzi, S. (ms), "Diaphonic Pluralism: How to be a Pluralist about Disagreement", unpublished manuscript.

Richard, M. 2012, "Reply to MacFarlane, Scharp, Shapiro, and Wright", *Philosophical Studies*, 160, 3, 477-95.

Richard, M. 2015, "What is Disagreement?", in *Truth and Truth-Bearers*, Oxford: Oxford University Press, 82-114.

Shah, N. and Velleman, D. 2005, "Doxastic Deliberation", *The Philosophical Review*, 114, 4, 497-534.

Shapiro, S. and Taschek, W. 1996, "Intuitionism, Pluralism, and Cognitive Command", *Journal of Philosophy*, 93, 74-88.

Shapiro, S. 2012, "Objectivity, Explanation, and Cognitive Command", in Coliva 2012, 211-37.

Smithies, D. 2012, "The Normative Role of Knowledge", *Noûs*, 46, 2, 265-88.

Wedgwood, R. 2007, *The Nature of Normativity,* Oxford: Oxford University Press.

Weiner, M. 2005, "Must We Know What We Say?", *The Philosophical Review*, 114, 227-51.

Williamson, T. 2000, *Knowledge and its Limits*, Oxford: Oxford University Press.

Wright, C. 1992, *Truth and Objectivity*, Cambridge, MA: Harvard University Press.

Wright, C. 2001, "On Being in a Quandary: Relativism, Vagueness, Logical Revisionism", *Mind*, 110, 437, 45-97.

Wright, C. 2006, "Intuitionism, Realism, Relativism and Rhubarb", in Greenough, P. and Lynch, M. (eds.), *Truth and Realism*, Oxford: Clarendon Press, 38-59.

Wright, C. 2012, "Replies Part III: Truth, Objectivity, Realism and Relativism", in Coliva 2012, 418-50.

Wittgenstein on Truth

Paul Horwich

New York University

Abstract

The topic is Wittgenstein's eventual abandonment of his *Tractatus* idea that a sentence is true if and only if *it depicts a possible fact that obtains*, and his coming (in the *Investigations*) to replace this with a *deflationary* view of truth. Three objection to the initial idea that will be discussed here are: (i) that its theory of 'depiction' relies on an unexplicated concept of word-object reference; (ii) that its notion of a possible fact *obtaining* (or existing, or being actual, or agreeing with reality) is also left mysterious; and (iii) that Wittgenstein's conception of *possible atomic facts* makes it difficult to see how any of them could fail to be actual. These problems are resolved by deflationism. But that perspective could not have been incorporated into the *Tractatus*. For the view of 'meaning qua use', on which deflationism depends, was the key insight enabling Wittgenstein to appreciate the untenability of his other central Tractarian doctrines.

Keywords: correspondence, deflationism, fact, *Investigations,* logic, picture theory, possibility, propositions, reference, Russell, *Tractatus*, truth, Wittgenstein

1. Introduction

This paper will address four related questions: What is the account of truth that Wittgenstein gives in the *Tractatus Logico-Philosophicus*?[1] To which view of the concept does he turn in his *Philosophical Investigations*?[2] Is this a move in the right direction? And how does it relate to other important differences between his early and late philosophy: is it a cause of them, a mere effect of them, or fairly independent of them?

Before getting started on all this, let me be upfront about something that will anyway become evident very quickly. I am a philosopher, but not much of a scholar. I am primarily interested in philosophical ideas, in the relationships between them, and in their plausibility—and less interested in whether they can be pinned on this or that philosopher at this or that point in their life. So my main concern in relation to Wittgenstein is not to decide what *exactly* he *meant*

[1] Wittgenstein 1922.
[2] Wittgenstein 1953.

Argumenta 2, 1 (2016): 95-105
ISSN 2465-2334

DOI 10.14275/2465-2334/20163.HOR

in his various writings. It is rather to examine and develop the material of substantial philosophical value that can be found in the *vicinity* of what he wrote.

In just this regard I greatly admire Saul Kripke's little book on Wittgenstein's 'private language argument'.[3] Not for its scholarship (for I think the ideas presented there are fairly far from anything in Wittgenstein himself), and not for the correctness of the philosophy (for I believe that those ideas are in themselves quite questionable); but rather for what Kripke aims to do, and succeeds in doing—which is to devise a line of thought that is inspired by Wittgenstein's writings and that, whether Wittgenstein's or not, and whether correct or not, deserves our attention. This is just the spirit in which I would like my own work on Wittgenstein to be taken, including the present paper.

2. The *Tractatus* View of Truth

What seems to jump out of the first few pages of the *Tractatus* is something like a *correspondence* theory of truth. But I say "something like", because a couple of qualifications must be made. In order to explain them, a few preliminaries are needed.

First: keep in mind that although we can speak of *sentences* (such as "snow is white") as "true" or "false", we more often apply those terms to *the things expressed by sentences* (such as the *proposition* that snow is white)—the objects of belief and assertion. Philosophers dispute which of these two ways of speaking is the more fundamental one; but most of us would agree that they are not the same; so distinct accounts of what we mean by them are called for.

Second: it is not easy to bring this distinction to bear on Wittgenstein's remarks about truth in the *Tractatus*. For his terminology sometimes diverges considerably from what is typically employed nowadays. In particular:

- He uses the term, "propositional sign" for what we might call an "uninterpreted sentence" (that is, a "sentence conceived of as a sequence of mere noises or inscriptions").
- He uses "proposition" (translated from the German, "*Satz*") for what we might say is a "significant sentence" or, in other words, an "uninterpreted sentence together with an interpretation of it, conceived of as something along the lines of 'the *reference-potentials* assigned to its component noises or inscriptions'".
- And he uses "the *sense* of a proposition" for what we might call "the *possible fact* that is represented by a significant sentence", or perhaps "the *Russellian proposition* (composed of objects and properties) that is expressed by a significant sentence".[4]

[3] Kripke 1992.

[4] A puzzling matter, that I shall simply flag and not attempt to resolve, is what Wittgenstein thinks must be added to a mere propositional sign in order to arrive at one of his *propositions.* He insists (3.13) that a *proposition* (in his usage of the term) does not contain its *sense* (in his usage). So the material to be added to the sign is *not* the possible fact it represents. But what else is available for that material to be? I have just hazarded "the reference-potentials of the words, qua noises, etc.". But what on earth is a reference-potential? Wittgenstein himself speaks of "feelers" that emanate from the components of a propositional sign and reach out to the referents of those components (2.1515). So a proposition, for him, is the propositional sign plus its 'feelers'—but *not* including the enti-

Translating his terminology into ours, it is fairly clear from the text that his overt theory of truth concerns significant *sentences*, and not what they express, or represent as being the case. His view (as we would put it) is that

A sentence is true iff
(i) it represents a certain possible fact; and
(ii) that possible fact is *actual*

which is tantamount to

S is a true sentence
≡ S represents a *fact* (= an *actualized possible fact*)

To see why this might aptly be termed a *correspondence* theory of truth we must look at Wittgenstein's distinctive account of *representation*.

His basic idea is that we should answer the relatively hard question of how a *sentence* (—a string of significant signs—) is able to represent something (—*that such-and-such is the case*—) by beginning with the relatively easy question of how a realistic picture, or a map, or an architect's model, represents what it does, and then proceeding to show that, initial appearances to the contrary, sentences represent in exactly the same way. Sentences are pictures!

More specifically, his view is that (i) a pictorial representation consists of elements arranged with respect to one another in a certain way; (ii) each such element has a referent; and (iii) the *actual* fact that the pictorial elements are arranged as they are represents the *possible* fact that the referents of those elements are also arranged just in that way.

For example: consider a map of Italy that has shaped dots, "★", "◆", and "■", whose referents are (respectively) Pisa, Rome, and Naples. The fact that, on the map, "◆" is between "★" and "■", *depicts* the possible fact that Rome is between Pisa and Naples. In this case, the common arrangement of pictorial elements and their referents is **that *x is between y and z***.

Moreover, according to Wittgenstein, we can and should regard a sentence (e.g. "John loves Mary") as a kind of pictorial fact (—that "John" is just to the left of "loves" which is just to the left of "Mary"). In this example, we are to suppose that:

- This fact has three referring elements—namely, the words, "John" and "Mary", and the relation, *x is just to the left of "loves" which is just to the left of y*.[5]
- The pictorial arrangement of these elements is not spatial, but is given by the abstract *logical* structure, **that *#(x,y)*.** This form is what the representing fact and the represented possible fact have in common.
- The referents of the three elements are, respectively, John, Mary, and the relation, *x loves y*.

ties they 'touch'. As far as I can tell, no *literal* account of this phenomenon is ever supplied.

[5] Note Wittgenstein's 3.1432: – We must not say, "The complex sign '*aRb*' says '*a* stands in relation *R* to *b*'"; but we must say, "*That* '*a*' stands in a certain relation to *'b*' says *that aRb*".

In our example, the "certain relation" is: *x is just to the left of "loves" which is just to the left of y.* The sentence's exemplification of this relation refers to an exemplification of the worldly relation, *x loves y*.

- Therefore, the depicted (represented) possible fact is just what we pre-theoretically know it to be—namely, the possibility *that John loves Mary*.

This is Wittgenstein's famous 'picture theory of meaning'. Clearly, *depiction of a possible fact* is treated as a form of *correspondence* to it. So it would seem that we *can* aptly say that Wittgenstein is proposing a "correspondence theory of truth".

But now let me elaborate the pair of reservations, to which I alluded at the outset, about the applicability of that label.

One of them is that, arguably, the above-sketched picture theory of sentential representation is supposed by Wittgenstein to apply *only* to what he calls "elementary propositions": that is, sentences that do not contain any logical vocabulary (either explicitly or implicitly).[6]For, if this is right, then Wittgenstein's correspondence theory of truth is also restricted to elementary sentences. Further principles will have to be added in order to extend that limited account into one that can cover logically complex sentences too. And such principles are indeed supplied in the *Tractatus*: they are the rules implicit in Wittgenstein's *truth-tables*—rules which specify how the truth or falsity of negations, disjunctions, conjunctions, and so on are determined by the truth and falsity of the elementary sentences to which the logical terms have been applied. Thus, what Wittgenstein really proposes is a *two*-stage theory of sentential truth, only the first of which invokes *correspondence*.

A second misgiving one might well have about calling Wittgenstein's view a "correspondence" theory of truth emerges from reflection on what he has to say about the other brand of truth I mentioned at the start: truth, not for *sentences,* but for what they express—that is, for what we nowadays call *propositions*.

One conclusion we might reach is that Wittgenstein does not have, and cannot have, *any* theory of truth of that sort. That is because the only kind of thing he countenances that resembles what are now called "propositions" are *possible facts*, and possible facts are not the sorts of things that it makes sense to speak of as "true or false". Rather, such things can only be "actual or non-actual" (that is, "actual or *merely* possible").

Alternatively, we might be inclined to think that, for us, "The proposition *that k is f* is true" and "The possible fact *that k is f* is actual" are just two ways of saying the same thing. In other words, we might suppose that Wittgensteinian

[6] Why this restriction? According to Wittgenstein, "My fundamental thought is that the 'logical constants' do not represent" (4.0312). That is to say, the words, "and", "not", "or", and so on, do not stand for bits of reality. And, in that case, how could his picture theory of representation conceivably work for sentences containing those words? Consider, for example, "It is raining or snowing". If his theory were to explain how this logically complex sentence represents what it does, the word "or" would either have to be part of a representing component of the sentence, or else it would have to be part of the pictorial arrangement of those components. But Wittgenstein's "fundamental thought" appears to preclude the first of these options. And his requirement that the represented entities exhibit exactly the *same* arrangement with respect to one another as the representing components appears to preclude the second. (Since surely the *noise* "or" does not feature in the possible fact that it is raining or snowing!)

Admittedly, Wittgenstein's formulations often seem to allow that his picture theory applies across the board. But in some passages (e.g. 4.0311) the restriction to sentences is explicit. And we have just seen why that is called for.

possible facts should be identified with what we call "propositions", so that they *can* perfectly well be spoken of either as "actual or non-actual" or as "true or false". In which case we would conclude that Wittgenstein is, after all, *implicitly* committed to a view of the notion that we call "propositional truth"— the view of it that coincides with what he *explicitly* says about *possible-fact actuality*.

I myself am not sure which of these answers is best (although I am inclined towards the second). For, on the one hand, I really do not see how, for example, "Massimo thinks *that Mars is green*" could be *ambiguous*: in one sense relating Massimo to a Russellian *proposition* and in another sense to a *possible fact*. But, on the other hand it is indeed hard to accept that what is said to be "true" could with equal propriety be described as "actual", and vice versa.

But whichever choice is made here, we can see that the *core* of Wittgenstein's proposal about *sentential* truth is the idea that only when the things meant (or *expressed*, or *represented*) have a certain special quality (—*being actual,* or *being factual*, or *obtaining*, or *existing*, or *agreeing with reality*, …—) can the sentences with those meanings be true. Thus one might well suppose that the character of this special quality will be crucial in judging whether his overall view can justly be called "a correspondence theory of truth".

But there appears to be no role for *correspondence* at this fundamental stage. The only point at which that notion enters Wittgenstein's picture is in his theory of *representation*—in the relationship between sentences and possible facts. We get to sentential *truth* only by relying on the concept of '*actuality'* (or '*obtaining',* or …) which Wittgenstein does not explain. So we might reasonably conclude that what we are given is *not* really a correspondence theory of *truth* but rather a limited correspondence theory of *representation* plus a primitivist *non*-theory of when the represented entities are facts.[7]

3. The *Investigations* View of Truth

Wittgenstein's remarks on truth in his much later work, *Philosophical Investigations,* suggest a position that is very different from the one to be found in the *Tractatus*.

Pretty clearly he is pivoting to a perspective that these days would be classified as "deflationary"—the term applied to accounts emphasizing that:

- Truth has *no* traditional explicit definition or reductive analysis (e.g. in terms of *correspondence*, or *coherence*, or *verifiability*, or *utility*, or *consensus*).

[7] Hans-Johann Glock puts the point nicely as follows: "The *Tractatus* marries a correspondence theory of depiction to an obtainment theory of truth" (Glock 2004). He goes on to allow that it is still appropriate to call Wittgenstein's view "a correspondence theory" since the similar view of truth that is proposed by both Moore and Russell around 1912 is standardly regarded as *paradigmatic* of such theories.

But it is worth noting a relevant *dis*-similarity. As we have seen, the foundation of Wittgenstein's account is his distinction between those possible facts that are *actual* and those that are not; and *actuality* is not a *correspondence* notion. But there is no analogous distinction in the account offered by Russell and Moore. They hold, for reasons we will be examining in section 4, that there can be no such things as *false propositions*. And this reasoning would lead them to the same skeptical conclusion about *merely possible facts.*

- Instead, the nature of the concept is implicitly fixed by the way that each statement specifies its own condition for being true—e.g. the statement *that lying is wrong* is true if and only if lying is wrong.
- It is an extremely *superficial* concept. There are hardly any concepts that are defined in terms of TRUTH, or whose possession requires prior possession of the concept, TRUTH.
- It is merely a useful *expressive* device, enabling certain generalizations to be formulated—for example, "All propositions of the form, <p or not-p>, are true", and "A belief is correct if and only if it is true".[8]

There is a lot of evidence in favor of attributing some such perspective to the later Wittgenstein.

First: we have the *Investigations,* section 136: "p is true = p". He is claiming here that ascribing truth to a proposition is equivalent to asserting the proposition itself.

Second: he was aware of, and influenced by, Frank Ramsey's advocacy of precisely that view.[9] They were together at Cambridge University in the early 1930s, and in the Preface of the *Investigations* he credits Ramsey with having been an enormous influence on his thinking.

Third: in accord with the deflationary definition and expressive *raison d'être* of TRUTH, this concept is given no important role in the *Investigations.*

And fourth: going hand-in-hand with deflationism about truth is the idea that our notions of predicative and nominal REFERENCE are fixed by the schemata:

f( ) is true of x = f(x)
n refers to x = n is x

which specify what a given concept applies to, *but only given prior possession of that concept* (which is deployed on the right-hand side of the relevant equations)—a possession that, on pain of circularity, cannot derive from knowing the concept's reference but must instead consist in mastery of its *use.* Thus Wittgenstein's move from the Tractarian *referential* conception of word-meanings to the Investigations *use* conception permits him to adopt a deflationary view of *reference*—and that is exactly what one would expect from a deflationist about *truth.*

Assuming it is right that the *Investigations* view of truth is deflationary, one might wonder which one of the various brands of deflationism on the market these days Wittgenstein favored, or would have favored. Would it be *disquotationalism*, according to which sentences (rather than propositions) are the bearers of truth, and the schema, "p" is true ↔ p, is the core of what implicitly defines the truth predicate? Or *pro-sententialism*, which denies that "true" is a genuine (logical) predicate, and which stresses instead the analogy between a pronoun and the sentence-type, "That is true" (insofar as both inherit their content from another expression (—one that is contextually salient)? Or the *redundancy theory*, whereby "The proposition *that p* is true" means exactly the same as just "p"? Or the *sentence-variable analysis*, which analyzes truth-talk in terms of quantification into sentence positions—"x is true" is taken to mean "(∃p)(x=<p> & p)"? Or *Tarski's theory*, which explains the truth condition of each sentence of a language

[8] For elaboration of these points see "What is truth?", Chapter 1 of Horwich 2010.
[9] Ramsey 1927.

in terms of the referential properties of its component words (characterized disquotationally) and the logical structure in which the words are embedded? Or *minimalism*, which resembles disquotationalism, except that it takes propositions (rather than sentences) to be the fundamental bearers of truth, and according to which our possession of the concept TRUTH is said to consist in our inclination to accept instances of the schema, "The proposition *that p* is true ↔ p"?

We do not have enough evidence to decide which of these (if any) he would prefer. But my guess is that Wittgenstein's aversion to philosophical theorizing would push him away from those versions of deflationism that come with a substantial amount of theoretical baggage, making them not *fully* deflationary.

So the redundancy theory would be disliked for its artificial and implausible conception of propositional identity (whereby, x can be the same proposition as y even though x involves the concepts of TRUTH and of PROPOSITION, and y does not).

Disquotationalism would be disliked for its misguided naturalistic presupposition that *propositions* are too weird to exist, and for its resulting revisionist disrespect for our *actual* use of "true"—our normal application of it to *the things we believe and disbelieve*, rather than to our *expressions* of such attitudes.

Tarski's theory would be disliked even more. For not only does it involve the same mis-motivated abhorrence of propositions and resulting focus on the truth of linguistic expressions—but, in addition, it repeat the Tractarian assumption (which he came to regard as deeply mistaken) that any meaningful sentence is either an elementary sentence or else is equivalent to the result of logical operations on elementary sentences.

The sentence-variable analysis would be disliked for its different form of revisionism. It defines a concept of 'truth' using concepts of variable and quantifier that we do not currently deploy. So the concept of truth that they help define cannot possibly be ours.

And the pro-sentential theory would be disliked for its scientistic overstretching of the analogy between our use of "That's true" to avoid repeating someone's recent assertion and, for instance, our use of "she" to avoid repeating some recent use of a name or description (e.g. "the discoverer of radium"). The most valuable of our deployments of "true" (as in "Some truths are unverifiable" and "Goldbach's conjecture is true") cannot be assimilated to that paradigm without absurd contortions.

Which leaves me thinking he would go for minimalism. (What a surprise!)

4. Which of Wittgenstein's Two Views of Truth is Better?

Well my opinion is already obvious. But let me supply some justification for it.

As indicated above, Wittgenstein's account in the *Tractatus* explains truth in terms of an unexplained (but equally puzzling) notion of *fact*. We are not told what it means to say "It's a *fact* that p", although this idea is obviously no less mysterious (and no less in need of explanation) than "It is *true* that p", which is what it is supposed to explain. This is a considerable defect. And, of course it will not help to define a "fact" as a possibility that "is actualized" or "obtains" or "exists" or "agrees with reality" unless we have explanations of at least one of these—which we do not.

Another difficulty with Wittgenstein's early position is that it is hard to see how there could be any difference between *actual* facts and merely *possible* facts.

After all, whether it is an actual fact *that Mars is* green or merely a possible fact that *Mars is* green, we are going to have the same constituents—namely, the object Mars and the property of being green—embedded in the same logical structure! Or to put the point in a slightly different way: the fact is a certain arrangement of certain entities; but the merely possible fact is exactly that arrangement of exactly those entities. So how could there exist any *merely possible facts*? Would not all *possible* facts have to be *actual*?

It is extremely surprising that Wittgenstein did not address this second problem, since his teacher and mentor, Bertrand Russell, made such a fuss about it. In his 1912 book-chapter, "Truth and Falsehood", and in subsequent writings, Russell argued in just this way that there could be no such things as *false* propositions.[10] He concluded that one cannot regard beliefs as relations between people and propositions (since many beliefs are mistaken). And so he devised a novel account of belief in which, for example, Massimo's belief *that Mars is green* is not a two-place relation between Massimo and the proposition (or possible fact) *that Mars is green*, but is rather a three-place relation between Massimo, Mars, and greenness. Russell showed this work to Wittgenstein, who dismissed the new theory of belief (on the grounds that it "didn't prohibit believing nonsense"). And Russell was demoralized, confiding to his then girlfriend that although he did not really understand the objection, his respect for Wittgenstein's insight made him feel that it must be right.[11] Still—and this is my main point—it is surprising that, on the one hand, Wittgenstein did not complain about Russell's argument for the non-existence of false propositions but, on the other hand, did not see that this argument would count equally against his own commitment to non-actual possible facts.

These two defects in the *Tractatus* account of truth are related to one another, as follows. In response to the first one it might be protested that Wittgenstein *does* address what it is for a possible atomic fact to be *actual* (or to *exist*, or to *obtain*,…). He says that this will be so when the constituents of the possible fact are *combined*. And this might seem to provide an illuminating analysis. But that is an illusion. The objection to it is not that *nothing* is said about what it is for objects to be *combined*. For Wittgenstein tells us that "In the atomic fact objects hang one in another, like the links of a chain" (2.03). Nor is the objection that this mere metaphor is woefully insufficient. For he makes it clear that the mode of combination is literally *logical*: it is for example, that Mars and being green (pretending for a second that these are basic entities) are embedded in the structure, 'that #(x)', or 'that x exemplifies #-ness'. The real objection is the *second* of the defects just sketched, which is that this answer fails to do what it was mainly supposed to do—namely, to distinguish those possible facts that obtain from those that do not. For, in both cases the same objects are embedded in the same structure. Thus we are left with no clue as to what it is for a possible fact to obtain, hence what it is for the sentence representing that possible fact to be true.

And a third objectionable feature of Wittgenstein's early theory is that the picture theory of representation, on which his view of sentential truth depends, deploys a unexplained relation of *reference* between words and things—a notion that is in no less need of definition than the notion of truth it is being used to define.

[10] Chapter 13 of Russell 1912.
[11] This biographical material is taken from pp. 80-82 of Monk 1990.

All of these related defects in the *Tractatus* account of truth—its reliance on unexplicated concepts of *actuality* and *reference*, and its failure to explain how *merely* possible facts could exist—can be removed in one fell deflationary swoop.

To begin with the problem of '*merely* possible facts'—let us call it "Russell's problem"—Wittgenstein can solve his own version of it by giving up the idea that there must be some *intrinsic* quality of a possible fact that makes it actual, and instead specifying conditions of *actuality* by means of the schema:

The possible fact *that p* is actual *iff* p.

And, of course, the parallel reply to Russell's argument against false propositions is that we can identify *the facts* with *the propositions that are true,* and then supply the conditions for propositions to be true by means of the schema:

<p> is true *iff* p.[12]

Again, the mistake was to think that there must be something in the *intrinsic nature* of a fact that distinguishes it from a mere proposition.

This stone also kills the second bird. What I mean is that in resolving the Russellian objection to Wittgenstein's early theory we are also addressing the objection that Wittgenstein's attempts to demystify TRUTH by explaining it in terms of FACT, which is equally mysterious. For, as we have seen, the solution again is to say:

It is a fact *that p iff* p.

Thirdly, regarding his reliance on an unexplained relation of reference: Wittgenstein's coming to appreciate that the meaning of a word is *not* constituted by its reference, but rather by how the word is used, allows him to explicate and demystify reference via a pair of schemata along the lines of

"n" refers to x ≡ n is identical to x (for names)

and

"f" is true of x ≡ f(x) (for predicates).

That is because our understanding of their right-hand sides will *not* already presuppose knowledge of the facts of reference that the left-hand sides are supposed to specify.

The overall moral is that the *Tractatus* theory of sentential truth needs an injection of deflationism if it is to be saved.

5. The Role Played by Wittgenstein's Accounts of Truth Within his Earlier and Later Philosophies

Now we might well wonder whether it is really possible to coherently inject the deflationary component of Wittgenstein's *Investigations* philosophy into his *Tractatus* philosophy.

[12] This is not the so-called *Identity Theory of Truth*, which is advocated by Jennifer Hornsby (cf. Hornsby 1997). Yes, she too maintains that true propositions are facts. But her idea is that this equation will define "true" in terms of "fact"—where the latter, roughly in the spirit of the *Tractatus*, is taken either to be a primitive or to be defined as a "combination of objects and properties". But my deflationary suggestion goes in the opposite explanatory direction—that we rely on the trivial Equivalence Schema to fix the concept of TRUTH, and then proceed to define "fact" as "true proposition".

I would suggest that the answer is no. That is basically because, as we have seen, deflationism about truth goes hand-in-hand with deflationism about reference; and deflationism about reference goes hand-in-hand with a use-theoretic view of meaning. But Wittgenstein's later commitment to 'meaning as use' is the principal basis for his recognizing his earlier "grave mistakes".[13] These included his failure to see:

- The fundamentally *instrumental* purpose of language
- The limitless variability of the functions of the different words in a language, and of the kinds of rules that must govern the uses of different words in order for them to perform this variety of functions
- The pervasiveness of vagueness and other forms of indeterminacy
- The absence of a small set of primitive words in terms of which all others can be defined
- The philosophical irrelevance of logic: in particular, that it does not provide the structure of language and thought
- The impossibility of drawing a line in advance around everything a language can be used to say
- The *real* source of philosophical confusion and pseudo problems—which is *not* the enormous conceptual distance between ordinary language expressions and their fundamental analyses (but rather our scientistic inclination to overstretch linguistic analogies)
- The contradiction between, on the one hand, the *theoretical* nature of his commitments (—to primitive terms, to the ideal of determinacy, to the fundamentality of logic, and to the exclusive role of *final conceptual analysis* in dissolving philosophical problems—) and, on the other hand, his bottom-line view that *philosophy cannot be theoretical*.

In short, once we let the genie out of the bottle—that is, the deflationary and use-theoretic cat out of its bag (the *Investigations*)—then the fundamental assumptions of his early philosophical system become evidently untenable.

So my conclusion is that Wittgenstein's magnum opus, *Philosophical Investigations,* is a great advance on the *Tractatus*, his brash, brilliant, *initial* attempt at a radically anti-theoretical philosophy. And this progress is almost entirely due to the way in which his new view of truth improves on the old one.[14]

References

Glock, H.-J. 2004, "Wittgenstein on Truth", in Löffler, W. and Weingartner, P. (eds.), *Knowledge and Belief*, Vienna: Hölder-Pichler-Tempsky, 328-46.

Hornsby, J. 1997, "Truth: The Identity Theory", *Proceedings of the Aristotelian Society*, 97, 1-24.

[13] See the Preface to *Philosophical Investigations.*

[14] My thanks to Hanjo Glock for the stimulus of his excellent above-cited article on this topic, to Charles Djordjevic for a most stimulating conversation about these matters, to Massimo Dell'Utri for his astute comments on the penultimate version of the present paper, and to all those who helped me by raising tough questions during my discussions of this material at the University of Zurich (June 2016), the University of Sassari, Sardinia (September 2016), and the University of Minnesota (November 2016).

Kripke, S. 1992, *Wittgenstein on Rules and Private Language*, Blackwell: Oxford.

Monk, R. 1990, *Ludwig Wittgenstein: The Duty of Genius*, New York: The Free Press.

Horwich, P.G. 2010, *Truth-Meaning-Reality*, Oxford: Clarendon Press.

Ramsey, F. 1927, "Facts and Propositions", in Mellor, D.H. (ed.), *Foundations: Essays in Philosophy, Logic, Mathematics and Economics*, London: Routledge and Kegan Paul, 1978, 40-57.

Russell, B. 1912, *The Problems of Philosophy,* New York: H. Holt and Company.

Wittgenstein, L. 1922, *Tractatus Logico-Philosophicus*, London: Kegan Paul (translated by C.K. Ogden).

Wittgenstein, L. 1953, *Philosophical Investigations*, Oxford: Blackwell (translated by G.E.M. Anscombe and G.H. von Wright).

Russellian Diagonal Arguments and Other Logico-Mathematical Tools in Metaphysics

Laureano Luna

I.E.S. Doctor Francisco Marín, Siles, Spain

Abstract

In its most general form, a diagonal argument is an argument intending to show that not all objects of a certain class C are in a certain set S, and does so by constructing a diagonal object, that is to say, an object of the class C so defined as to be other than all the objects in S. We revise three arguments inspired by the Russell paradox (an argument against Computationalism, an argument against Physicalism, and a counterargument to the Platonic One Over Many argument), extract its underlying structure, and suggest a criterion to tell the ones that end up at a paradoxical object like the old Russell set from the ones that could actually accomplish a diagonalization. We conclude with the suggestion that the use of logico-mathematical tools, which is a significant methodological contribution of the analytical tradition, opens up a promising line of research in metaphysics.

Keywords: Analytical Philosophy, Computationalism, Physicalism, Platonic Forms, Sets, Diagonalization, Indefinite Extensibility, Russell's Paradox, Axiom of Replacement, Metaphysics.

1. Four Russellian Diagonal Arguments in Metaphysics

In its most general form, a diagonal argument is an argument that shows that not all objects of a certain class C are in a certain set S and does so by constructing (usually by reference to S) a diagonal object, that is to say, an object of class C that is other than all the objects in S. We expound three arguments concerning metaphysics, all of them inspired by the Russell paradox, extract its underlying structure, and suggest a criterion to tell the ones that end up at a paradoxical object like the old Russell set from the ones that could actually be able to deliver a diagonal object.

Luna and Small (2009) have put forward a Russellian diagonal argument against Computationalism. Computationalism is the thesis that all thought acts are computations (executions of algorithms) so that there is a correspondence between thought types and algorithms. Below is a version of the argument. By 'thought' we will mean hereafter 'thought type'. Note that we define a special re-

Argumenta 2, 1 (2016): 107-123
ISSN 2465-2334

DOI 10.14275/2465-2334/20163.LUN

lation and denote it by marking 'about' with an asterisk: 'about*'; this relation is not exactly the one Luna and Small use but it will do the job as well.

ARGUMENT 1

Assume Computationalism.
If Computationalism is true, there is a function f from algorithms to thoughts, representing the correspondence between the former and the latter, such that, for any thought t, there is an algorithm g such that f(g)=t.
Let us say that a thought is about* x if and only if ('iff', henceforth) it asserts a proposition of the form '$\forall y\ (y \in S_1 \rightarrow y \in S_2)$', for some sets S_1, S_2, and $x \in S_1$.
Call an algorithm g normal iff f(g) exists and is not about* g; let S be the set of all normal algorithms.
Let t^* be a thought asserting just '$\forall x\ (x \in S \rightarrow x \in S)$' and let $t^*=f(g^*)$.
t^* is exactly about* all normal algorithms.
Then, by the usual Russellian reasoning,[1] g^* is normal iff it isn't. Contradiction.
Therefore, f does not exist and Computationalism is not true. □

The aboutness* relation involved in the argument may look contrived but all that matters for the validity of the argument is that it be well-defined, and there is no obvious reason to believe it is not. The mention of sets S_1 and S_2 obeys the reason that quantifiers bounded by predicates like 'all P' are usually granted to successfully quantify over all P if P's extension is a set.

Argument 1 invites a parallel argument against Physicalism if by Physicalism we understand the claim that mental states or thoughts are so dependent on brain states ('brainstates', hereafter) that no thought exists without a corresponding brainstate and no two different thoughts can accompany one and the same brainstate-type: this is often called 'type supervenience Physicalism'. By 'brainstate' we will mean 'brainstate-type' hereafter.

ARGUMENT 2

Assume Physicalism.
If Physicalism is true, there is a function f from brainstates to thoughts such that, for any thought t, there is a brainstate b such that f(b)=t and t is the thought that accompanies b.
Let us say that a thought is about* x if and only iff it asserts a proposition of the form '$\forall y\ (y \in S_1 \rightarrow y \in S_2)$', for some sets S_1, S_2, and $x \in S_1$.
Call a brainstate b normal iff f(b) exists and is not about* b; let S be the set of all normal brainstates.
Let t^* be a thought asserting just '$\forall x\ (x \in S \rightarrow x \in S)$' and let $t^*=f(b^*)$.
t^* is exactly about* all normal brainstates.
Then b^* is normal iff it isn't. Contradiction.
Therefore, f does not exist and Physicalism is not true. □

[1] Assume g^* is normal; then t^* is about* g^* (for it is exactly about* all normal algorithms) and this makes g^* not normal. Assume g^* is not normal; then t^* is not about g^* and this makes g^* normal.

Nothing is evidently wrong in the argument. However, the fact that it can be easily parodied should cast the shadow of a doubt upon it. For consider the following reasoning, to which we have given an unnecessarily complex form to mirror the structure of arguments 1 and 2 (this is why some phrases are in parentheses):

ARGUMENT 3

Assume there are thoughts.
If there are thoughts, there is an identity function f from thoughts to thoughts such that, for any thought t, (there is a thought t such that) f(t)=t.
Let us say that a thought is about* x if and only iff it asserts a proposition of the form '$\forall y\ (y \in S_1 \rightarrow y \in S_2)$', for some sets S_1, S_2, and $x \in S_1$.
Call a thought t normal iff f(t) (exists and) is not about* t.
Let θ^* be a thought asserting just '$\forall x\ (x \in S \rightarrow x \in S)$' (and let $\theta^* = f(\theta^*)$).
θ^* is exactly about* all normal thoughts.
Then θ^* is normal iff it isn't. Contradiction.
Therefore, f does not exist and there are no thoughts. □

This reasoning has the same structure as Russell's paradox, of which the following is a version:

ARGUMENT 4

Assume there are sets.
If there are sets, there is an identity function f from sets to sets such that, for any set s, (there is a set s such that) f(s)=s.
Call a set s normal iff f(s) (exists and) $s \notin f(s)$.
Let s^* be the set of all normal sets (and let $s^* = f(s^*)$).
Then s^* is normal iff it isn't. Contradiction.
Therefore, f does not exist and there are no sets. □

Russell's famous paradox uses the set theoretical membership relation instead of the aboutness* relation above defined. Of course, arguments 3 and 4 can be easily simplified: function f in each of them serves the unique purpose to make apparent that they can be given the same structure as arguments 1 and 2.

The underlying structure of these Russellian diagonal arguments is a *reductio*:

1. We assume there is a surjective function f: $A \rightarrow B$.
2. We define a relation R relating members of B and members of A.
3. We define a member b^* of B having R to exactly all members of A not related by R to their images by f (i.e. to all *normal* members of A).
4. b^* is the image by f of some member a^* of A.
3. b^* has R to a^* iff it doesn't. Contradiction.
5. Therefore, f does not exist. □

The object b^* is the diagonal object: we use it to diagonalize out of the range of f, that is to say, to construct a member of B that is the image by f of no member of A, showing f is not surjective.

The structure of these diagonal arguments responds to Priest's Inclosure Schema (Priest 2002: 134). Priest defines a set $\Omega = \{x: \varphi(x)\}$, for some property

φ, and assumes ψ(Ω) for some property ψ; then, for each x⊆Ω which has property ψ, he defines a diagonalizing function δ such that δ(x)∉x and δ(x)∈Ω. Obviously, this schema leads to the contradiction that δ(Ω)∈Ω and δ(Ω)∉Ω. Priests is inclined to endorse the contradiction; if we are not, we can infer that if δ exists, then there is no set Ω of all φ-objects that has property ψ. In our pattern, property φ would be 'being a member of B' and property ψ would be 'being the image under f of some subset of A', and δ(x) would be the diagonal object b^* we produce by means of relation R. Thus, the arguments conclude that B is not the image under f of a subset of A. The situation could also be depicted in the terms of Shapiro and Wright (2006) by saying that property φ is *indefinitely extensible relative* to property ψ, which as before implies that there is no set of all φ-objects that has property ψ.[2] The contradiction arrived at on each occasion depends on a first order validity sometimes called 'Thomson's theorem' (Thomson 1962): ~∃x ∀y (Rxy ↔ ~Ryy).

Arguments 3 and 4 are obviously unsound but it is not obvious what is wrong with them.

2. Diagonal Arguments and Paradox

Let us first address argument 4, which is essentially the Russell paradox.

Certainly, a number of authors, when dealing with Russell's paradox, limit themselves to the conclusion that the diagonal object does not exist on pain of contradiction. Even if the conclusion is true, acknowledging its truth does not provide us with an explanation thereof, hence also not with a solution to the paradox. Simply stating that there is no set of all non self-membered sets because the extension of the concept of non self-membered set is too large to form a set is no explanation at all. If s^* (i.e. the old Russell's set) does not exist, there must be an explanation of why and how its definition fails to define a set. We will offer the two main narratives that aim at an explanation of the facts; we will call them the conventional (the term implies no pejorative connotation) and the alternative narrative.

Let us address the conventional approach in the first place. How the definition of s^* in argument 4 (namely, 'set of all normal sets') fails to define a totality is incomprehensible if normality of sets is well-defined. So, we should explore the possibility that normality is not well-defined. But for normality to be ill-defined, set membership must be ill-defined. And in fact, we usually admit it is ill-defined in a sense, namely, in the sense that there is no definite totality (i.e. no set) of all sets. The multiplicity of sets does not make up a definite totality; it is *indefinitely extensible* or *open ended:* there are sets beyond any set of sets.[3] Hence, sets are not given all at once but come in stages or levels; one cannot have a definite totality of them all; rather, for any definite totality of sets one can define or think of, there are sets beyond that totality, at a higher level. Each time we de-

[2] I thank an anonymous reviewer for suggesting the convenience of mentioning Priest's Inclosure Schema in this context. We are using a simplified version of Shapiro and Wright's *relative indefinite extensibility.*

[3] See Russell 1905 for an early discussion of the topic, though an expression translatable by 'indefinite extensibility' have I found nowhere before Zermelo's 1930 "schrankenlose Fortsetzbarkeit".

fine a totality of sets, we rise to some level of a hierarchy of such totalities, with more sets showing up at further levels.[4]

Accordingly, there must be levels of membership and levels of normality. But the definition of s^* does not specify any level of normality. In failing to distinguish levels of normality, the definition of s^* may be overlooking a necessary hierarchy and committing vicious circularity: after all, it would be defining membership in s^* in terms of self-membership of all sets, s^* included (we will dwell below on the circularity in the definition of s^* and the ensuing necessity to distinguish levels of set membership and normality). These reasons would explain why the definition of the diagonal object fails and the object itself does not exist. Such is the conventional narrative in its bare bones.

There is an alternative account. Some logicians believe our quantifiers can only range over objects that are previously available, so that they never range over indefinitely extensible multiplicities but only over set-sized portions of them that are, if not otherwise, determined by context.[5] If this is actually so, the level of normality of s^* is implicit in the definition of s^* and determined by context in such a way as to avoid circularity. s^* just stands at a higher level than all the sets its definition is about. The definition of s^* would refer to normal sets that are normal at a level of normality that is below the level of normality at which s^* could be or fail to be normal. Then, s^* could be normal without containing itself because it would possess normality at a higher level than its members.[6] This reading of the definition of s^* dissolves (rather than solves) the paradox: the appearance of a paradox would arise from an incorrect reading of the definition of the diagonal object. In this interpretation, s^* is not at all paradoxical; it just diagonalizes out of the sets its definition is about, extending the universe of discourse of its definition by one set, namely, s^*. This alternative approach is typically preferred by those logicians who believe that absolutely unrestricted quantification is impossible and the universe of discourse is always restricted to a set-sized multiplicity, which may be indefinitely extended by diagonalization (for an overview of positions concerning the possibility of unrestricted quantification, see Rayo and Uzquiano 2006, Introduction).[7]

What about argument 3? Is there a related way out? It seems so, since, arguably, there is no set of all thoughts. Patrick Grim (1991, ch. IV) has displayed a number of arguments to show that there is no set of all truths and some of them can easily be transferred from truths to thoughts. One of them is just a Russellian diagonalization argument (Grim 1991: 110-13), of which this is a version:

[4] This is of course the *iterative conception* of sets, according to which sets come in stages forming an indefinitely extensible hierarchy in which sets are always posterior in the hierarchy to their members.

[5] This theory is congenial to the usual model theoretic principle that all universes of discourse are sets.

[6] Note that s^*, if it exists, is normal at some level of normality; otherwise, it would be a member of itself, and only normal sets are members of s^*.

[7] The alternative approach faces this objection: why can normality-at-any-level not be used to reproduce the paradox? But the answer seems straightforward: 'normality-at-any-level' fails to quantify over all levels because the hierarchy of levels is indefinitely extensible. The alternative account does little more than assuming that the hierarchy proposed by the conventional account is in fact always implicitly present in our discourse.

> Let T be any thinkable (i.e. definable, constructible or susceptible of specification) set of thoughts; define some aboutness* relation from thoughts in T to any objects; construct a diagonal thought that is exactly about* all thoughts in T not about* themselves; if the diagonal thought were in T, it would be about* itself iff it were not; hence, it is not in T; rather, it diagonalizes out of T; as the diagonalization procedure can be achieved for any thinkable set of thoughts, there is no thinkable set of all thoughts; but if the set of all thoughts existed, it would be thinkable, since it could be thought of as the set of all thoughts; thus, it does not exist.[8]

Here is a related argument based on an idea by Russell 1903 (par. 500: 538-39) that employs no aboutness relation:

> For any definable set of thoughts x, let $\pi(x)$ be its product, which is the thought that all thoughts in x are thoughts. For any such x, $\pi(x)$ exists (even if $x=\varnothing$ and $\pi(x)$ is vacuously true). Let s be any definable set of thoughts. Let R be the (possibly empty) set of all products $\pi(x)$ in s such that $\pi(x)\notin x$. R is definable because s is; hence, $\pi(R)$ exists. Assume $\pi(R)\in s$. Then $\pi(R)\in R$ iff $\pi(R)\notin R$, which is a contradiction. Hence, $\pi(R)\notin s$ and there is a thought not in s. As s was any definable set of thoughts, there is no definable set of all thoughts. But if the set of all thoughts existed, it would be definable as the set of all thoughts. Therefore, the set of all thoughts does not exist.[9]

This is exactly how Russell's paradox is customarily used to prove there is no set of all sets: for each set s of sets, there is a set not in s, namely, the set of all non self-membered sets in s. In these arguments we apply the diagonalization procedure only to sets and we assume that the diagonal objects cannot be ill-defined; indeed, as we deal only with multiplicities that are sets, we can no longer argue that the aboutness or membership relations are ill-defined because thoughts or sets should come in levels: those thoughts or sets that can be put in a set can be given all at once because they do make up a definite totality.

So, we can escape argument 3 in the same way we avoided argument 4. In the conventional narrative, since it is incomprehensible how t^* could fail to exist if normality of thoughts is well-defined, we are compelled to assume that normality is not well-defined; and it is not because the multiplicity of thoughts is not a definite totality; it is indefinitely extensible: there are thoughts beyond any set of thoughts. Thoughts come in levels. Accordingly, there must be levels of aboutness* and levels of normality; then θ^*, the purported thought about* all normal thoughts, is defined with vicious circularity because its definition does not specify a particular level of normality.

In the alternative narrative, the level of normality is implicit in the definition of θ^*, and θ^* is about* normal thoughts that are normal at a level of normality that is below the level of normality at which θ^* could be or fail to be normal. If so, θ^* can be normal without being one of the thoughts θ^* is about and no paradox exists: θ^* simply diagonalizes out of the thoughts it is about*.

[8] See also Luna and Small 2009: 88-89.

[9] One may be tempted to think that the definable objects must form a set because there are at most a countably infinite number of definitions; however, Richard's paradox strongly suggests that natural language is indefinitely extensible. Luna and Taylor 2010 propose that some syntactical expressions must define different objects in different logical contexts due to inevitable ambiguity in the range of quantifiers. See also Luna 2013.

3. Arguments 1 and 2 vs. Arguments 3 and 4

Let us reckon what the fate of arguments 1 and 2 would be if we could apply to them what the conventional or what the alternative narrative has to say about Russell's paradox. In the conventional approach, we would accuse the diagonal objects of being ill-defined and we would declare them nonexistent, which would indeed refute the arguments. In the alternative narrative, we would claim that the diagonal objects are not comprehensive enough in their scopes to bring about the contradictions they seem to provoke (for they would fail to be about* all normal objects) and this would all the same refute the arguments. Obviously, if one of these criticisms is not applicable because the argument's framework is not of the proper kind, neither is the other, since those criticisms are alternate treatments of one and the same type of situation, namely, multiplicities that can never be entirely given.

So let us first examine whether we can level against the diagonal objects in arguments 1 and 2 the same type of counterargument the conventional narrative employs against the existence of s^* and θ^*. Can the diagonal objects in these arguments be ill-defined for the same reason as s^* and θ^* are in the conventional narrative?

As regards argument 1, Luna and Small deny that possibility. They argue that, at least under the Church-Turing thesis, the set of all algorithms not only exists, it is effectively enumerable: it is the set of all Turing machines; hence, according to the model theoretical principle that any nonempty set is a possible domain of discourse, we should be able to refer to them all in order to diagonalize out of them; the authors assume that principle and call it the *principle of semantic clarity*. In the terms of our approach here, we would say that contrary to what happened with sets and thoughts, if algorithms can all be put in one and the same set, that is to say, if they are all given at once and need not come in levels, concerns regarding levels of normality and circularity in the definition of the diagonal object are out of place. One has to be a strict finitist, which is an extremely radical position to adopt in philosophy of mathematics, to deny the existence of the set of all Turing machines. And even if one rejects the Church-Turing thesis and believes that algorithms exist that cannot be represented by Turing machines, one will most probably believe that algorithms, whatever they are, do form a set. Thus, as the central claim of Luna and Small seems plausible, one has to avow that there does seem to be a difference between this case and arguments 3 and 4.

If we approach argument 2 in this spirit, the relevant question is whether there is a set of all possible brainstates. We have no better reason to believe that there is no such set than we have to believe there is no set of all possible (types of) earthquakes or (types of) atoms, for instance. Brainstates are much the same kind of objects as types of earthquakes or as elements in the periodic table: they are types of physical objects and these are not the kind of objects we expect to constitute indefinitely extensible multiplicities. Usually, if a class C of objects is indefinitely extensible, it is the case that we can employ the definition of an arbitrary set of C-objects to define a new C-object not in the set, so diagonalizing out of the set. But possible types of physical objects, like types of earthquakes, do seem to be possibly given all at once because their givenness appears to be absolutely independent of our definitions of them: we can hardly make new types of earthquakes emerge by diagonalizing once and again out of sets of types of

earthquakes into an indefinitely extensible hierarchy of levels. Be it as it may, if the set of all possible brainstates does exist, brainstates and normality need not come in levels, and the proposition stated by thought t^* is, for all we know, well-defined; if so, there is no evident reason to deny that t^* is a possible thought that diagonalizes out of all thoughts in the range of f. One must grant at least that there does *seem* to be a relevant difference between t^*—in arguments 1 and 2—and the diagonal objects s^* and θ^* that, according to tradition, are paradoxical.

If the set of all algorithms and the set of all brainstates actually exist, the alternative approach to Russell's paradox has nothing to say about arguments 1 and 2, since it only deals with cases involving indefinitely extensible multiplicities.

Let us take this as a preliminary approach to our subject. In order to gain additional insight, we need to examine in some detail the topic of circularity in the definition of s^* in argument 4. Laurence Goldstein (2009), reasoning within the conventional narrative, believed that it is possible not just to prove the non-existence of s^* by *reductio* but also to explain and render it intuitive by showing why its definition fails to define a set. He pinpointed circularity in the definition of s^* in the following way.

Goldstein points out that if s^* existed, its definition would fail to define a set because it would be viciously circular; and it would be viciously circular because the expression that defines s^*:

$$\forall x\,(x \in s^* \leftrightarrow x \notin x)$$

can be developed into the infinite conjunction of one sentence of the following form for each set s:

$$s \in s^* \leftrightarrow s \notin s.$$

If s^* existed, one of these sentences would be

$$s^* \in s^* \leftrightarrow s^* \notin s^*$$

which would render the definition of s^* clearly circular, besides inconsistent. Since s^* is specified by its definition, it can only exist if its definition succeeds in defining a set. But if s^* exists, its definition fails and s^* does not exist; as a consequence, s^* does not exist.

Goldstein's idea suggests the necessity of distinguishing levels of set membership in order to avoid viciously circular definitions. This in turn suggests another approach to the problem of the circularity in s^*. Note that normality is defined upon the set membership relation. For normality to be well-defined, the set membership relation should be determinate when normality is defined upon it. But that is not the case if s^* exists because membership in s^* depends on normality, for s^* is defined to have precisely all normal sets as members. So, if s^* exists, normality is defined upon set membership and set membership (in s^*) is defined upon normality. In order to disentangle our definitions, we should distinguish alternate levels of membership and normality: membership_0 is membership before any normality has been defined; normality_0 is normality defined upon membership_0; membership_1 is membership after normality_0 has been defined; normality_1 is normality defined upon membership_1, etc.

One can indeed obtain a set of all normal sets at each level of normality but one can extend the levels through all ordinals, which go beyond sethood. Hence, the necessity of distinguishing levels of normality implies that the objects for which normality is defined do not make up a set but an indefinitely extensible

multiplicity. *If the objects for which normality is defined do form a set, then the hierarchy of normality levels is not indefinitely extensible because there is a level at which all normality levels are already available, namely, the level at which those objects form a set; this should permit to use a definition of normality simultaneously valid for all levels, if such levels exist at all.*

It is easy to see how the alternative narrative would have it here; it would contend that, since normality can only be defined upon what is already determinate, it is *in fact* not defined by reference to membership in s^* itself; so, s^* stands at a higher level of normality than all the normal sets its definition is about. This happens, so to say, in an automatic way. Hence, the definition of s^* cannot but diagonalize out of all the normal sets it refers to, and reading it otherwise makes no sense. Obviously, this approach only applies when the objects for which normality is defined do not form a set, for if they do, there is a highest level that contains all levels and out of which it is impossible to diagonalize. Both analyses of argument 4, that of the conventional and that of the alternative approach, are easily applied to argument 3 after the suitable replacements; essentially, one must substitute thoughts for sets and the aboutness* relation for the set membership relation.

In the following paragraphs, we will deal simultaneously with arguments 1 and 2 though explicitly referring solely to argument 1: applying to argument 2 what we will say about argument 1 is straightforward.[10]

The question is whether normality and the diagonal object in argument 1 are circularly defined for the same reason as they are in argument 3. t^*, the diagonal object in argument 3, is exactly about* all normal algorithms but if g^*—such that $f(g^*)=t^*$—exists, then aboutness* seems ill-defined for t^* because it is defined through these clauses: for each algorithm g,

$$t^* \text{ about}^* g \leftrightarrow f(g) \text{ not about}^* g$$

among which

$$t^* \text{ about}^* g^* \leftrightarrow t^* \text{ not about}^* g^*.$$

It is clear that if there is a g^* such that $f(g^*)=t^*$, then normality, as it occurs in the definition of t^*, is ill-defined. This seems to leave us with a disjunction: *either* normality in the definition of t^* is well-defined and g^* does not exist *or* normality in the definition of t^* is ill-defined (so that t^* does not exist and the existence of g^* is not even an issue). But this disjunction proves nothing. One can choose the first disjunct if one wishes to save the argument or the second if one prefers to keep Computationalism. So, to rescue the argument, the second disjunct must be shown false or, at least, it must be shown that it is false if Computationalism is true. We will argue for the latter, that is, we will argue that Computationalism would imply that normality in the definition of t^* is well-defined.

It would seem that argument 1 contains the same vicious circularity as arguments 3 and 4: in argument 1, normality is defined upon the aboutness* relation; hence, for normality to be well-defined, this relation should be determinate before we define normality; but it seems it is not, because aboutness* is defined by means of the expression '$\forall y\ (y \in S_1 \rightarrow y \in S_2)$' where S_1 can be the set of all normal algorithms, as it is in fact assumed to be for t^*. Thus, at least if t^* exists,

[10] In order to apply to argument 2 what refers to argument 1 in the following paragraphs, just replace 'argument 1' by 'argument 2',' algorithm' by 'brainstate', 'g^*' by 'b^*', 'g' by 'b', 'Computationalism' by 'Physicalism', and 'syntactical' by 'physical'.

normality involves aboutness* and aboutness* involves normality. So, it appears that to avoid circularity, we should introduce alternate levels of aboutness* and normality: aboutness*$_0$ is aboutness* before any normality is defined; normality$_0$ is normality defined upon aboutness*$_0$; aboutness*$_1$ is aboutness* after normality$_0$ has been defined; normality$_1$ is normality defined upon aboutness*$_1$, etc.

But this hierarchy of levels would take us too far if Computationalism is true and all thoughts are algorithms; it would take us beyond sethood, which is impossible if algorithms form a set, as they seem to do. If the set of all algorithms exists, normality cannot come in an indefinitely extensible hierarchy of levels: there must be a level at which normality of algorithms is fully determinate (namely. the level at which the set of all algorithms becomes available) and at that level we can profit from that determinateness to carry out the diagonalization procedure successfully. This situation makes it dubious that we can escape argument 1 by applying the same strategies we used against arguments 3 and 4: the fact that, in all evidence, algorithms do form a set would stay in our way.

As regards the determinateness which we claim algorithms possess and which should render normality well-defined in argument 1, consider that algorithms are syntactical in nature and syntactical facts are always determinate: Gödel showed they are equivalent to arithmetical facts. It is precisely this difference between semantical and syntactical properties that makes the whole difference between Gödel's famous self-referential sentence G and the Liar. G only involves the syntactical property of provability within a formal system and this makes of it a definite mathematical statement that cannot be paradoxical.[11] That syntactical facts (and most probably physical facts too) are determinate in a sense in which others are not can also be illustrated by this example: note that one can easily produce a paradox by referring to semantical properties of thoughts or of propositions as in 'what I am now thinking is false' or 'this proposition is not true' but it is hard to figure out how one could produce a paradox if one refers only to syntactical features as in 'this sentence has five words' or only to physical properties as in 'my current brainstate involves at least one billion synapses'. There is surely a relation between determinateness understood as the property of adding up to a definite totality and need not come in (an indefinitely extensible hierarchy of) levels and determinateness understood as the property of being so definite as to preclude paradox. This relation, however, requires further research to be developed elsewhere.

If Computationalism is true, the circularity in the definition of normality in argument 1 can only be apparent. If f exists and thoughts are linked by f to algorithms, any property or relation of thoughts is as determinate as syntactical facts are. Furthermore, if algorithms make up a set, as they seem to do, they need not come in levels; they may be given all at once. Confessedly, the case for the determinateness of physical facts is less conclusive than the case for the determinateness of syntactical facts. However, it would be really strange if brainstates were unable to form a set, for the objects in indefinitely extensible classes seem to depend on our capability to construct new objects by diagonalization and

[11] As an anonymous reviewer points out, there are indeed proofs of (versions of) Gödel's theorem that involve semantical notions (such as 'truth' or 'model'). However, the relevant fact is that the self-referential G only involves the syntactical predicate of formal provability: this is why it cannot be paradoxical. The nature of the proof of G's undecidability is irrelevant for this purpose.

physical states of affairs do not appear to be the kind of things whose existence would depend on our constructions.

The response to the criticism of argument 1 in the alternative account would be as follows: there cannot be normal algorithms at higher levels of normality than t^* can be about, because algorithms form a set, so that all of them can be simultaneously given, making up a possible universe of discourse.

4. Arguments 1 and 2 in Set Theoretical Terms: A Soundness Criterion

So far, our analysis has revealed that thoughts do not seem able to make up a set whereas algorithms do.[12] This permits to recast argument 1 in set theoretical terms. If thoughts do not form a set and algorithms do, then there can be no such surjective function as f from algorithms to thoughts: the set theoretic axiom of Replacement would prohibit its existence. The axiom of Replacement states that, if the domain of a function is a set, its range is a set too. Therefore, if algorithms form a set and f exists, then also thoughts form a set, which seems implausible. So, on the plausible assumption that there is a set of all algorithms but no set of all thoughts, we can use the axiom to argue that f does not exist and Computationalism is false. This is not the place to revise and discuss the justifications of the axiom of Replacement. We will only remark that it is so widely accepted, be it for its mathematical fruitfulness or its philosophical plausibility, that showing it incompatible with the thesis that each thought corresponds to some algorithm is enough to make a case against Computationalism.[13]

The fate of a Russellian diagonal argument seems to depend on whether the members of the multiplicity for which normality is defined form a definite totality and can be given all at once or they form an indefinitely extensible class and come in levels. If argument 1 succeeds as a diagonal argument, its success depends crucially on these facts:

1. There is a set of all algorithms but there is no set of all thoughts: thoughts form an indefinitely extensible multiplicity.
2. If Computationalism is right and f exists, then by the set theoretical axiom of Replacement, there is a set s_f of thoughts that is the range of f; but since thoughts form an indefinitely extensible multiplicity, it is possible to diagonalize out of any definable set of thoughts, hence also out of s_f; so, we can produce a thought that is not in s_f but diagonalizes out of it, thereby proving Computationalism false.

Consider this simplified argumental blueprint: there is no set of all thoughts but, since a set of all algorithms does exist, if Computationalism were right and f existed, Replacement would imply the existence of the set of all thoughts; thus, Computationalism is wrong. This fact would on its own refute Computationalism as defined but it would not refute a weaker form of Computationalism, namely, the thesis that all possible *human* thoughts are so related to algorithms

[12] An anonymous reviewer reminds me of the fact that some axiomatics (e.g. NBG) admit classes too big to be sets, usually called *proper classes;* so another way to express the difference would be: the class of all thoughts is a proper class while the class of all algorithms is a set.

[13] A *locus classicus* for the discussion of the rationale of the axiom is chapter 10 of Parsons 1983.

that f exists with respect to them; it would not, because computationalists could easily raise the counterargument that those thoughts that are not in the range of f, even if they are possible in some abstract sense, may not be possible *human* thoughts; that is, they might be impossible for creatures whose thoughts are linked through f to algorithms. But argument 1 goes a decisive step further and constructs one diagonal thought—hence one possible *human* thought—that is not in the range of f; thus, the argument, if successful, refutes also that weaker form of Computationalism.

The weaker form of Computationalism can also be argued against by means of a slight modification of the argumental blueprint above: there is no set of all possible *human* thoughts but, since a set of all algorithms does exist, if Computationalism were right and f existed, Replacement would imply the existence of the set of all possible *human* thoughts; thus, Computationalism is wrong. The problem is that so far, we have argued against the existence of a set of all possible thoughts but not against a set of all possible *human* thoughts. Here is an argument against the existence of such a set, framed along the lines of the Russellian argument on page 5:

> For any definable set of possible thoughts x, let $\pi(x)$ be its product, i.e. the thought that all thoughts in x are thoughts. For any such x, $\pi(x)$ exists as a possible human thought (because x is definable), even if $x=\varnothing$ and $\pi(x)$ is vacuously true. Let s be any definable set of possible human thoughts. Let R be the (possibly empty) set of all products $\pi(x)$ in s such that $\pi(x)\notin x$. R is definable because s is; hence, $\pi(R)$ exists as a possible human thought. Assume $\pi(R)\in s$. Then $\pi(R)\in R$ iff $\pi(R)\notin R$, which is a contradiction. By *reductio*, $\pi(R)\notin s$ and there is a possible human thought that is not in s. As s was any definable set of possible human thoughts, there is no definable set of all possible human thoughts. But if the set of all possible human thoughts existed, it would be definable as the set of all possible human thoughts. Therefore, the set of all possible human thoughts does not exist.[14]

When one compares the Luna-Small Russellian diagonal argument with Russell's paradox and its conventional solution—or more generally arguments 1 and 2, on the one hand, with arguments 3 and 4, on the other—a difference becomes apparent. As regards the latter, we have, at least in the conventional narrative, a standard reason to believe that the diagonal object is not well-defined, namely, that its definition fails to distinguish among the different levels of a property (i.e. normality) and this failure makes the definition fail as such. But this is not the case for the former: if algorithms or brainstates make up a definite totality (as we have good reasons to believe), then they can be all given at once and a definite aboutness* relation together with a one-level normality property must exist for them.

So, the criterion for distinguishing a Russellian construction that leads to a paradoxical object (in the conventional account) or fails to produce the desired contradiction (in the alternative one) from a genuine diagonal argument is this: do the objects for which normality is defined form a definite totality that can be given once for all or are they members of an indefinitely extensible hierarchy and come in levels?

[14] Recall that the analysis of argument 1 accomplished from footnote 10 to this can be rendered an analysis of argument 2 by making the substitutions described in footnote 10.

If they come in levels, normality without a level index may be ill-defined for them (which would cast just as much doubt on the existence of the diagonal object as there is about the existence of the old Russell set) or it may fail to be inclusive enough to bring about a contradiction. Otherwise, the case is essentially other than Russell's paradox.

As a consequence, unless we can substantiate the claim that algorithms or brainstates are unable to form definite totalities, we should not dismiss arguments 1 and 2 on the grounds that they are but avatars of the old Russellian paradox. *It is noteworthy that neither computationalists nor physicalists have seriously addressed the problem that human thoughts appear to spread along a hierarchy of levels that extend beyond sethood whereas algorithms and brainstates seem to be given or available once for all, so as to make up definite totalities.* Upon analysis, this is ultimately the state of affairs that renders arguments 1 and 2 plausible.

5. Assessing a Russellian Argument against Platonic Forms

Let us finally consider a Russellian argument proposed by Michael Loux (1998: 34-35)[15] though forms of the argument appear already in Russell (1903, par. 78: 80-81) and in Mally (1914: 225). It is ultimately an intensional version of Russell's set theoretical paradox. Loux' purpose is to deny that a famous Platonic thesis—to be found in *The Republic*, Book X, 596a-b[16]—is tenable in full generality. The thesis contends that whenever many different things are of a same kind, so that there is a same name convening to all of them, a corresponding Form exists (this is sometimes called the One Over Many argument for the existence of Platonic Forms). So, for example, a beautiful poem and a beautiful melody have beauty in common and we say that they are both beautiful; hence, according to the thesis, beauty exists as a Form. Thus, the argument turns the property P common to all entities in some collection c into a Platonic Form F_P. Thus, the following argument assumes, for *reductio,* that, for any collection c_P containing all the objects that have some property P in common, there is a Platonic Form F_P corresponding to P.

ARGUMENT 5

Assume the One Over Many argument.

As the One Over Many argument is right, there is a function f that takes a collection c_P of all objects sharing some property P and returns the corresponding Platonic Form F_P.

Call a Platonic Form F_P normal iff $F_P \notin c_P$, that is, iff it does not exemplify itself. For instance, the Platonic Form T corresponding to the property of being a table is normal because T is not a table but a Platonic Form.

Being a normal Platonic Form is a property which T has in common with other Platonic Forms (e.g. being a chair). Let c^* be the collection of them all. Then $f(c^*) = F^*$ exists.

F^* is the Form corresponding to the property of being a normal Form.

But F* is normal iff it isn't. Contradiction.

Therefore, f does not exist and the One Over Many argument is unsound. □

[15] I am indebted to James Grindeland for bringing this argument to my attention.

[16] See for instance Plato 1992: 265.

We have good reasons to believe that, even if Platonic Forms exist, there is no set of them all. We can argue, for instance, that there is at least one Platonic Form for each set s, namely, the Platonic Form corresponding to the property of being a member of s.[17] If so, Platonic Forms come in levels and normality for them is ill-defined in the sense in which normality is ill-defined for sets in argument 4 and for thoughts in argument 3. Therefore, either F^* is ill-defined and does not exist (as the conventional approach would have it) or, if it does exist, then it can be normal without contradiction at a higher level than all Forms exemplifying it, as the alternative approach to paradoxes would contend.

6. Mathematical Tools in Metaphysical Argumentation

Arguments 1 and 2 belong in a family of anti-reductionistic arguments attempting to show that thoughts are too different from other kinds of objects to be ontologically reduced to them or to be put in some (too narrow) dependence relations with them. For instance, it has been argued that thoughts have an intrinsic semantic nature while algorithms do not or that brainstates are spatial but thoughts are not. The novelty is that the difference here invoked is ultimately of mathematical nature: adding up to a definite totality (so algorithms and brainstates) vs. being spread along an indefinitely extensible hierarchy of levels.

The phenomenologist Ernst Mally (1914) used an argument very similar to argument 3 to support the claim that 'thought D is directed toward thought D' is meaningless because a duly typed language—akin to the proposed by Russell and Whitehead in the *Principia*—would not allow for it. If that sentence were not meaningless—Mally argues—it would make sense as well to construct a thought G directed exactly to all thoughts that are not directed to themselves; and G would be paradoxical.

There is a well-known argument against physicalism due to Kripke (Kripke 1980). Kripke relies on the following proof that true identities are necessary:

1. $\forall P \forall xy\ (x=y \rightarrow (Px \rightarrow Py))$	Premise (indistinguishability of identicals)
2. $\forall x\ (\Box(x=x))$	Premise (necessity of self-identity)
3. $\forall xy\ (x=y \rightarrow (\Box(x=x) \rightarrow \Box(x=y)))$	1, Universal Instantiation for 'P'
4. $\forall xy\ (x=y \rightarrow \Box(x=y))$	2, 3, Propositional logic.

Kripke adds a conceivability argument: the identity of mental and physical states is at most contingent since its falsity is conceivable; and concludes that the identity is false. No matter how controversial, it is an example of a formal argument in metaphysics.

A set theoretical argument against *global supervenience materialism* (the thesis that the whole sphere of the mental supervenes on the whole physical state of the world) has been proposed by Franz von Kutschera (1994). The following is a version of that argument.

If we represent each proposition as the set of all possible worlds at which it is true, there are more possibly true propositions (i.e. propositions that are true at some possible world) than possible worlds, because of Cantor's theorem. This suggests that not all possibly true propositions can be believed (the original argument elaborates on this point). Let us divide the class of all possible proposi-

[17] Except the empty set, if we reject necessarily uninstantiated forms.

tions into two (not *a priori* exclusive) classes: *doxastic propositions,* which are the propositions about states of belief such as 'that 1+1=2 is believed', and *objective propositions,* which are the propositions made true or false by the state of the physical world. It seems that all possible objective propositions can be believed (unfortunately, the original argument does not elaborate on this). As a consequence, not all doxastic propositions are objective. But if states of belief supervened upon physical states, all doxastic propositions would be objective propositions. Therefore, the targeted kind of supervenience must fail. There are a number of difficulties with this argument but, at the very least, it must be credited the audacity of suggesting that mental states and physical states may make up multiplicities with different mathematical properties.

Michael Detlefsen has published a paper (Detlefsen 2002) in which he utilizes Löb's theorem (Löb 1954) to reveal some difficulties of Computationalism (called *mechanism* by the author). Löb's theorem states that for any consistent arithmetical system Σ and any formula φ, if Σ proves 'if φ is Σ-provable, then φ', then φ is Σ-provable.[18] Detlefsen's most significant conclusion in the referenced paper is that, on plausible assumptions, our proof resources cannot regard themselves both as reliable and as mechanizable. Assume they consider themselves both things, reliable and mechanizable. As they believe they are mechanizable, they regard themselves as subject to Löb's theorem. As they believe they are reliable, they prove for any sentence *s* 'if *s* is provable, then *s*' but they cannot consider themselves able to prove all sentences, because they could hardly consider themselves reliable if they believed they prove a sentence and its negation; but this is obviously incompatible with being subject to Löb's theorem; so they cannot consider themselves subject to such theorem; and we have reached a contradiction: they regard themselves as subject to Löb's theorem and they do not. As far as I can see, the argument has no evident flaw.

Arguments such as Mally's, Kripke's, Kutschera's, Detlefsen's, the original Luna-Small argument or its extension in this paper are likely to look suspect to a number of readers who may distrust metaphysical arguments based on logico-mathematical phenomena, regarding them as extremely likely to contain some fallacious sleight of hand. In some cases these suspicions may rest on a prejudicial belief in the existence of some iron curtain isolating the realm of the metaphysical from the realm of the logico-mathematical. This prejudice may be one in a bundle of inherited beliefs evidencing just inertial resistance to disappear. In the last decades, the analytical tradition has passed from a plain rejection of metaphysics (inspired by neo-positivist empiricism) to a cautious approach to some metaphysical issues, and in this transition it has brought with itself the use of logico-mathematical tools for the treatment of philosophical topics, hence also of metaphysical ones. This can be seen as one of the most significant contributions of the analytical tradition to contemporary metaphysics.

As far as I know, argumentation from logical phenomena to properly metaphysical topics was inaugurated by Mally in 1914 with the argument sketched above. Gödel's famous Gibbs lecture some sixty years ago (Gödel 1951) is a perspicuous case of this type of argumentation. Lucas' and Penrose's Gödelian arguments against computationalism (Lucas 1961, Penrose 1989, 1994), even if not equally esteemed by all, have spurred prolific discussion over decades. Alvin

[18] Löb's theorem relies on a standard form of representing the provability in Σ in Σ's language.

Plantinga (Plantinga 1974) is the most notorious of several philosophers who have used some features of the Kripkean accessibility relation among possible worlds (Kripke 1963) to argue for the existence of God. Patrick Grim (Grim 1988, 1991) has harnessed some topics in set and model theory for theological purposes, namely, to set up arguments against the possibility of divine omniscience. Arguments of this kind have compelled some theists to espouse a sort of *process theology,* in which God is not immutable, so modifying the traditional definition of God.[19]

One of the goals of this paper is to make apparent that this type of argumentation could be a promising line of metaphysical research even if to date it seems suspect to many and is for the most part absent from mainstream metaphysical discussion.

References

Brendel, E. 2001, "Allwissenheit und 'offenes Philosophieren'", *Erkenntnis*, 54 (1), 7-16.

Butler, R.J. (ed.) 1962, *Analytical Philosophy*, I, New York: Barnes and Noble.

Detlefsen, M. 2002, "Löb's Theorem as a Limitation on Mechanism", *Minds and Machines*, 12 (3), 353-81.

Gödel, K. 1951, "Some Basic Theorems on the Foundations of Mathematics and their Philosophical Implications", in *Collected Works*, III, *Unpublished Essays and Lectures*, Feferman, S. *et al.* (eds.), Oxford: Oxford University Press, 1995, 304-23.

Goldstein, L. 2009, "A Consistent Way with Paradox", *Philosophical Studies*, 144 (3), 377-89.

Grim, P. 1988, "Logic and Limits of Knowledge and Truth", *Nous*, 22, 341-67; reprinted in Martin, M. and Monnier, R. (eds.), *The Impossibility of God*, Amherst, New York: Prometheus Books, 2003, 381-407.

Grim, P. 1991, The Incomplete Universe: Totality, Knowledge, and Truth, Cambridge, MA: The MIT Press.

Kripke, S. 1963, "Semantical Considerations on Modal Logic", *Acta Philosophica Fennica*, 16, 83-94.

Kripke, S.A. 1980, Naming and Necessity, Cambridge, MA: Harvard University Press.

Kutschera, F. von 1994, "Global Supervenience and Belief", *Journal of Philosophical Logic*, 23, 103-10.

Loux, M. 1998, A Contemporary Introduction to Metaphysics, New York: Routledge.

Löb, M.H. 1955, "Solution of a Problem of Leon Henkin", *The Journal of Symbolic Logic*, 20, 115-18.

Lucas, J.R. 1961, "Minds, Machines, and Gödel", *Philosophy*, XXXVI, 112-27.

Luna, L. and Small, C. 2009, "Intentionality and Computationalism. A Diagonal Argument", *Mind & Matter*, 7 (1), 81-90.

[19] This seems to be the case for Brendel 2001, though her reference is not just to Grim's but also, and fundamentally, to W.K. Essler's akin argumentation.

Luna, L. and Taylor, W. 2010, "Cantor's Proof in the Full Definable Universe", *Australasian Journal of Logic*, 9, 10-25.

Luna, L. 2013, "Indefinite Extensibility in Natural Language", *The Monist*, 96 (2), 295-308.

Mally, E. 1914, "On the Objects' Independence from Thought", *Zeitschrift für Philosophie und philosophische Kritik*, CLV, 1, 37-52; transl. by D. Jacquette in *Man and World*, 22, 215-31, 1989.

Parsons, C. 1983, *Mathematics in Philosophy*, Ithaca, New York: Cornell University Press.

Penrose, R. 1989, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford: Oxford University Press.

Penrose, R. 1994, *Shadows of the Mind. A Search for the Missing Science of Consciousness*, Oxford: Oxford University Press.

Plantinga, A. 1974, *The Nature of Necessity*, Oxford: Oxford University Press.

Plato, 1992, *The Republic*, translated by G.M.A. Grube, 2nd edition, Indianapolis, IN: Hackett Publishing Co.

Priest, G. 2002, *Beyond the Limits of Thought*, New York: Oxford University Press.

Rayo, A. and Uzquiano, G. (eds.) 2006, *Absolute Generality*, New York: Oxford University Press.

Russell, B. 1903, *Principles of Mathematics*, Cambridge: Cambridge University Press; reedited by Routledge, London-New York, 2010.

Russell, B. 1905, "On Some Difficulties in the Theory of Transfinite Numbers and Order Types", *Proceedings of the London Mathematical Society*, 2 (4), 29-53.

Thomson, J.F. 1962, "On some Paradoxes", in Butler 1962, 104-19.

Shapiro, S. and Wright, C. 2006, "All Things Indefinitely Extensible", in Rayo and Uzquiano 2006, 255-304.

Zermelo, E. 1930, "Über Grenzzahlen und Mengenbereiche. Neue Untersuchungen über die Grundlagen der Mengenlehre", *Fundamenta Mathematicae*, 16 (1), 29-47.

The Contemporary Relevance of Peirce's Views on the Logic and Metaphysics of Relations

Claudine Tiercelin

Collège de France

Abstract

Independently of Frege or Russell, C.S. Peirce made major contributions to the history of the logic and metaphysics of relations. After presenting his metaphysical interpretation of relations and his emphasis on the reality and irreducibility of relations, the paper shows how Peirce's views are tied to the dispositional realism he defends within a scientific realistic metaphysics, and why they are still relevant for assessing the logical and ontological status of relations, and insightful for the metaphysical agenda to pursue today.

Keywords: Logic of relations, Metaphysics of relations, Dispositional realism

1. Introduction: From Metaphysics to Logic and Vice Versa

For the great American metaphysician Charles Sanders Peirce, logic and metaphysics were going hand in hand. "Metaphysics consists in the result of the absolute acceptance of logical principles not merely as regulatively valid, but as truths of being" (1.487).[1] However, before becoming scientific and realistic, metaphysics had a first therapeutic duty: to make our ideas clear about what counts as a real or a pseudo metaphysical issue, and, in particular, about one's position on the problem of universals: should one side with the nominalists or with the realists? In that respect, Randall Dipert is right when he claims that "logic, especially the logic of relations played a central role in the development of Peirce's philosophy" (2004: 287) and that his logic of relations had a decisive impact on the right realistic metaphysics one should adopt:

> My plan for defeating nominalism is not simple nor direct; but it seems to me sure to be decisive and to afford no difficulties except the mathematical toil it requires. For as soon as you have once mounted the vantage ground of the logic of relatives,

[1] (1.487) refers to volume 1, paragraph 487 of Peirce (1931-58) (8 vols.). All references to Peirce will be to this edition.

Argumenta 2, 1 (2016): 125-138
ISSN 2465-2334

DOI 10.14275/2465-2334/20163.TIE

> which is related to ordinary logic precisely as the geometry of three dimensions of geometry of points on a line, as soon as you have scaled this height, I say, you find that you command the whole citadel of nominalism which must thereupon fall almost without another blow (4.1).

Indeed, Peirce thought that the Logic of Relatives had clearly shown that we needed to change our formulation of the problem of universals from "are universals real?" to "are continua real?". And to operate such a change, he made three main moves in his account of relations: he suggested a new definition of propositional form, stressed the existence of three necessary and sufficient categories, claimed the impossible reduction of triadic relations to dyadic ones.

I shall first remind of Peirce's main contributions, independently of Frege or Russell, to the history of the logic and metaphysics of relations at the end of the 19th century, before presenting his metaphysical interpretation of relations and the emphasis he put, within his categorial framework, on the reality and irreducibility of relations, and, even more, on the irreducibility of triadic relations. After explaining in what way Peirce's understanding of relations is part and parcel of the kind of dispositional realism he defends within an overall scientific realistic metaphysics, I shall claim that such a framework is still relevant both for evaluating the logical and ontological status of relations as such, and, more importantly, as a source of inspiration for the right metaphysical agenda to pursue today.

2. Peirce's Contribution to the History of the Logic of the Relations

Peirce's contributions to logical theory are numerous and profound. His work on relations, building on ideas of De Morgan, influenced Schröder and, through Schröder, Peano, Russell, Löwenheim and much of contemporary logical theory. Peirce had an extensive development of a symbolic relational logic. As has been underlined, although Frege anticipated much of Peirce's work on relations and quantification theory, and extended it more, Frege's work remained out of the mainstream until the twentieth century. Thus, it is plausible that Peirce's influence on the development of logic has been of the same order as Frege's (Tiercelin 1991; Dipert 1995). However, in contrast to Frege's highly systematic and thoroughly developed work in logic, Peirce's work remains fragmentary and extensive, rich with profound ideas, but most of them left in a rough and incomplete form (Michael 1974; Merrill 1978). Yet, it is possible to highlight some evolution and major influences of such ideas on Peirce.

Indeed, prior to his long *Description of A Notation for the Logic of Relatives, resulting from an Amplification of the Conceptions of Boole's Calculus* (3.45 ff), in fact prior to his published papers on Boolean algebra and syllogistic of 1867, Peirce had devoted some study to relational terms and to their role in arguments, which had led him, in the early 1860's, to see the incompleteness of traditional syllogistic and of Boole's algebra of classes, and the necessity of taking relations into account. Some of Peirce's discussion of relations is conducted in the context of the deduction of his categories, as spelled out in his now classical paper *On a New List of Categories* (1867), in which he distinguishes three main concepts: the ground, the relate and the correlate. Let us take a quality like "white": whiteness is the ground, the basis on which a thing is said to be white, that is, x is white on the ground that x embodies whiteness. In a quality like "greater", the greatness of a

thing by comparison with another is the ground (the basis) of attributing this quality to that thing. A ground is an abstraction, a Form (e.g. greatness, whiteness) that is the basis for attributing a quality to things. Wherever two things are brought into relation, one of them is taken as the relate (or the subject of the proposition), the other as the correlate (or the direct object of the proposition). For example, in "A kills B", A is the subject of the relation (the relate), B is the object of the relation (the correlate). Another important distinction (see Lowell Lectures 1866), concerns two basic kinds of relation: relations of concurrence (agreement, equiparance: "that of relates whose reference to a ground is a prescindable or internal quality") and relations of opposition (difference, disquiparance: "that of relates whose reference is an unprescindable or relative quality"). Any proposition involves one of these. The influence of scholastic views, and particularly of Occam on Peirce's early work on relations (whether through direct reading, or through Mill's System of Logic or Prantl) is more substantial than mere adaptation of terminology (Michael 1974: 48; Tiercelin 1991: 46-55; 188-193). Indeed, Peirce's initial distinction is close to Occam's distinction between connotative (monadic predicates like "white") and relative (dyadic predicates like "father") terms. Such terms are similar in so far as they do not directly refer to individual objects, but rather refer to such objects obliquely or indirectly. As such, they primarily signify a meaning and secondarily signify individual objects on the basis of that meaning. A term like "Socrates" refers to its object (Socrates) directly; a term like "white" refers to its objects (Socrates, Plato, etc.) indirectly through its meaning, namely something having begotten a son. In relative terms, reference to a direct object (the son of Socrates) is required by the meaning of the term itself. In the proposition "Socrates is white", white refers to Socrates on the ground of his having whiteness. In the proposition "Socrates is a father", "father" refers to Socrates on the ground of his having begotten a son. In any proposition that asserts, the character indicated by the predicate term is asserted of the object indicated by the subject term. As such, in a true proposition, the predicate term is said to include in its reference what the subject term indicates. "The same thing is meant by 'the stove is black' as by 'there is blackness in the stove'; embodying blackness is the equivalent of black" (1.551). Black refers to the stove on the ground of its embodying blackness. Hence, it refers primarily to its meaning and secondarily to objects on the basis of that meaning. In Peirce's view and in Occam's view, then, a connotative or relative term refers to objects on the basis of its reference to a meaning (Tiercelin 1993: 188-193).

Peirce's *Description of a Notation for the Logic of Relatives, resulting from an Amplification of the Conceptions of Boole's Calculus of Logic* (1870) is undoubtedly one of the most important works in the history of logic. It is in this paper that a notation for multiplying quantified relations and techniques for manipulating them first appear. A "relative" is viewed as a term in the sense in which it is used by the Aristotelian logicians, that is, the relationship between a relative and a relation. Hence, a relative term does "double duty" (Dipert 2004: 296), semantically representing a certain extension or class, namely the "logical sum" of ordered pairs (n-tuples) of individuals: this is precisely the modern semantic understanding of the extension of a relation of *n* places as a set of *n*-tuples. But it also serves as an operation on classes.

However, as Dipert has pointed out, two things should be noted. First of all, this paper of 1870 was not the very first symbolic treatment of relations: credit for this should go to Lambert, or, better known and crucially influential on Peirce, to

De Morgan (*On Syllogism IV*, 1859) to whom Peirce explicitly refers in 1866 and 1867, who fruitfully applied the concepts of Boolean algebra to relations (Thibaud 1975 and Martin 1980). And it is the same idea which Peirce applied in 1870 to what he named "relatives" or "relative terms". Secondly, the 1870 paper was less a rupture with the preceding framework as viewed by Peirce as an "enriched", "beautified" and "completed" *generalization* of it (4.5).

This said, "Peirce's 1870 paper is remarkable for its sheer imaginativeness, but also for its disorderly presentation" (Dipert 2004: 297). In many cases, the development amounts to experimentation with various notations for relations which he never used again, and to the following out of algebraic analogies (such as with exponentiation and a binomial theorem, something Boole too attempted, though not for relatives). However, the basic techniques allowed Peirce to express very complex quantified relational statements and often to show their equivalence to other statements. For example, whatever is lover of or servant to a woman is the same class as the non-relational logical addition of the lovers of a woman and the servants of a woman.

$$(l +, s) = l\text{w}+, s\text{w}$$

Here relations are indicated by italicized letters, and simple classes by non italicized letters. Juxtaposition indicates a notion of "application" of a relative to a class, and not any sort of ordinary logical multiplication (intersection of classes), showing how a relative behaves more like a function or operator than a class or term. It is an equivocation, however, often made by modern set theory, just as predicates were also conceived as "propositional functions" by Frege, Russell and Whitehead (Martin 1980; Dipert 2004: 296).

3. Three Main Logical and Metaphysical Results

The first result of such an analysis concerns the evolution from a *grammatical* to a *logical* approach of the structure of propositions. In 1867, the list of categories was derived from the functions or logical forms of *judgements*: although the subject predicate form was already greatly re-arranged, in particular, by means of the medieval tools provided by the theory of *suppositio* (Tiercelin 1993: 48-55). We now have a new way of characterizing a proposition and several original definitions of *rhemes*, *relatives*, *relationships*, and *relations*:

> An assertion fulfilling the condition having been obtained, let a number of the proper designations of individual subjects be omitted, so that the assertion becomes a mere blank form for an assertion which can be reconverted into an assertion by filling all the blanks with proper names. I term such a blank form a rheme (4.354).

In a complete proposition, there are no blanks, and it is called by Peirce a medad, or medadic relative:

> A non-relative name with a substantive verb, as '—is a man', or 'man that is—' or '—'s manhood' has one blank; it is a monad, or monadic relative. An ordinary relative with an active verb as '—is a lover of—' or 'the loving by— of—' has two blanks; it is a dyad, or dyadic relative. A higher relative similarly treated has a plurality of blanks. It may be called a polyad. The rank of a relative among these

> may be called its adinity, that is, the peculiar quality of the number it embodies (3. 465).

Hence a *relative* may be defined as "the equivalent of a word or phrase which, either as it is (when I term it a complete relative), or else when the verb 'is' is attached to it (and if it wants such attachment, I term it a nominal relative), becomes a sentence with some number of proper names left blank". A *relationship*, or *fundamentum relationis*, is said to be "a fact relative to a number of objects, considered apart from those objects, as if, after the statement of the fact, the designations of those objects had been erased". A *relation* is a relationship considered as something that may be said to be true of one of the objects, the others being separated from the relationship yet kept in view. Thus, for each relationship there are as many relations as there are blanks. For example, corresponding to the relationship which consists in one thing loving another, there are two relations, that of loving and that of being loved by. There is a nominal relative for each of these relations, as 'lover of—', and 'loved by—'. These nominal relatives belonging to one relationship are in their relation to one another termed correlatives. In the case of a dyad, the two correlatives, and the corresponding relations are said each to be the converse of the other. The objects whose designations fill the blanks of a complete relative are called the correlates. The correlate to which a nominal relative is attributed is called the relate". Indeed, a relation "is a fact about a number of things". Thus, the fact that a locomotive blows off steam constitutes a relation, or more accurately a relationship between the locomotive and the steam. We may go so far as to say that "in reality, every fact is a relation. Thus, that an object is blue consists of the peculiar regular action of that object on human eyes". And, Peirce claims, this is "what should be understood by the 'relativity of knowledge'" (3.416).

The second important result has to do with Peirce's emphasis on the fact that we need *three* in order to have a relation, hence a relation cannot be reduced to a mere "connexion" between two things:

> Is relation anything more than a connexion between two things? For example, can we not state that A gives B to C without using any other relational phrase than that one thing is connected with another? Let us try. We have the general idea of giving. Connected with it are the general ideas of giver, gift, and 'donnée'. We have also a particular transaction connected with no general idea except through that of giving. We have a first party connected with this transaction and also with the general idea of giver. We have a second party connected with that transaction, and also with the general idea of 'donnée'. We have a subject connected with that transaction and also with the general idea of gift. A is the only haecceity directly connected with the first party; C is the only haecceity directly connected with the second party, B is the only haecceity directly connected with the subject. Does not this long statement amount to this, that A gives B to C? (3.464).

Indeed, Peirce claims that "*in order to have a distinct conception of Relation, it is necessary not merely to answer this question but to comprehend the reason of the answer*" (italics mine) (3.464). Suppose you thought instead that relations were nothing but *connexions* of two things. Then "all things would be equally connected", and "nothing could be more connected with one idea than with another". Now, suppose you make "the relation of any two things consist in their connexion being connected with a general idea". Then, since "that last connexion is, on your own

principles, itself a relation, and you are thus defining relation by relation; and if for the second occurrence you substitute the definition, you have to repeat the substitution *ad infinitum*". And you will be "guilty of a *circulus in definiendo*" (3.464).

From such observations, Peirce concludes that three categories, in other words, tokens, icons and indices, are both *necessary* and *sufficient*:

> A dual relative term, such as "lover" or "servant", is a sort of blank form, where there are two places left blank. I mean that in building a sentence round "lover", as the principal word of the predicate, we are at liberty to make anything we see fit the subject, and then, besides that, anything we please the object of the action of loving. But a triple relative term such as "giver" has two correlates, and is thus a blank form with three places left blank. Consequently, we can take two of these triple relatives and fill up one blank place in each with the same letter, X, which has only the force of a pronoun or identifying index, and then the two taken together will form a whole having four blank places; and from that we can go on in a similar way to any higher number. But when we attempt to imitate this proceeding with dual relatives, and combine two of them by means of an X, we find we only have two blank places in the combination, just as we had in either of the relatives taken by itself. A road with only three-way forkings may have any number of termini, but no number of straight roads put end on end will give more than two termini. *Thus any number, however large, can be built out of triads; and consequently no idea can be involved in such a number, radically different from the idea of three* [italics mine]. I do not mean to deny that the higher numbers may present interesting special configurations from which notions may be derived of more or less general applicability; but these cannot rise to the height of philosophical categories so fundamental as those that have been considered (1.363).

Hence, a third crucial result. Not only are there "no more Kainopythagorean categories than these three. For the first category is non relative experience, the second is experience of a dyadic relation, and the third is experience of a triadic relation", but also and more importantly, "it is impossible to analyze a triadic relation, or fact about three objects, into dyadic relations; for the very idea of a compound supposes two parts, at least, and a whole, or three objects, at least, in all. On the other hand, every tetradic relation, or fact about four objects can be analyzed into a compound of triadic relations" (7.537).

Although Peirce's "remarkable theorem" of the irreducibility of triadic relations was later to be shown as false in terms of the modern logic, and may have to be related, as Dipert has rightly insisted on, both to Peirce's wish to favor a graphical system and to the influence of the chemical framework on many of his views, it undoubtedly contributed to underline the problem of what a logical form is and, in many respects, to come to right metaphysical results. So it is to these that we should now turn.

4. Relations, Dispositions, and Peirce's Metaphysical Defense of Dispositional Realism

Peirce called himself "an Aristotelian of the scholastic wing, approaching Scotism", or "a scholastic realist of a somewhat extreme stripe". By what he meant, first, that, contrary to what is often asserted today, when it comes to the realism/anti-realism issue about universals, the problem is not that of wondering

whether there *exist* universals apart from our ideas or words. For a scholastic realist, reality should not be equated with existence, which is but a mode of reality. Though what exists is real, what is real may not exist; existence is reaction, interaction—the characteristic mode of being of particulars, of seconds (Haack 1992: 22). Peirce thought that there was indeed a "nominalistic Platonism" (8.10) which consisted in conceiving the existence of things "independent of all relation to the mind's conception of it" (8.13), hence, in viewing universals like "man" or "horse" as referring to abstract particulars or existents. Now, scholastic realism should refuse to take universal or singular entities as *utterly* independent of thought and signification: "The real is that which is not whatever we may happen to think it, but is unaffected by what we may think of it." (8.12; 1871) "The real is that which *signifies* something real" (5.320). Hence:

> Anybody may happen to opine that 'the' is a real English word; but that will not constitute him a realist. But if he thinks that, whether the word 'hard' itself be real or not, the property, the character, the predicate, *hardness* is not invented by men, as the word is, but is really and truly in the hard things and is one in them all, as a description of habit, disposition, or behavior, *then*, he is a realist (1.27n1).

As a first consequence, individuals can be said to *exist*, but not, strictly speaking, to be *real*:

> We can only say, in a general way, that a term, however determinate, may be made more determinate still, but not that it can be made absolutely determinate. Such a term as 'the second Philip of Macedon' is still capable of logical division––into Philip drunk and Philip sober, for example; but we call it individual because that which is denoted by it is in only one place at one time. It is a term not absolutely indivisible, but indivisible as long as we neglect differences of time and the differences which accompany them. Such differences we habitually disregard in the logical division of substances. In the division of relations, etc., we do not, of course, disregard these differences, but we disregard some others (3.93).

In particular, as is shown by the logic of relatives, there are

> three kinds of terms which involve general suppositions of individual cases. The first are individual terms, which denote only individuals; the second are those relatives whose correlatives are individual: I term these infinitesimal relatives; the third are individual infinitesimal relatives, and these I term elementary relatives (3.95).

As a second consequence of such a realism and by means again of the logic of relatives, in saying that generals are real, Peirce claims, first, that generals do not so much apply to "classes" or "collections" than to "systems" (4.5), and, secondly, *continuity* being the real general, that one should subscribe to real modalities, *real possibilities* and *real necessities* (4.172):

> None of the scholastic logics fails to explain that *sol* is a general term; because although there happens to be but one sun, yet the term *sol aptum natum est dici de multis*. But that is most inadequately expressed. If *sol* is apt to be predicated of many, it is apt to be predicated of any multitude however great, and since there is no maximum multitude, those objects, of which it is fit to be predicated, form an

> aggregate that exceeds all multitude. Take any two possible objects that might be called suns and, however much alike they may be, any multitude whatsoever of intermediate suns are alternatively possible, and therefore, as before, these intermediate possible suns transcend all multitude. In short, the idea of a general involves the idea of possible variations which no multitude of existent things could exhaust but would leave between any two not merely many possibilities, but possibilities absolutely beyond all multitude (5.102).

As a third and important consequence of adopting such a "scholastic realism" of "real possibilities", which Peirce clearly intended as a piece of scientific metaphysics, we should start by securing the semantic level and, in particular, be clear about the claim that *there are real dispositions*, the meaning of our dispositional attributions, and the reasons why the reduction of dispositional ascriptions to conditionals does not seem to work (due, for example, to finkish or antidote cases), or why reduction sentences may or may not tell us "all" that dispositional predicates mean. Indeed, as such dispositional realists as Mellor, Ellis or Mumford insist on today, Peirce was convinced that *one should look for real dispositional properties and not mere predicates*, and that properties are not (or are not given) simply by the meaning of our predicates. In other words, we want a conditional and non truth-functional statement such as "if *x* was dropped, it would break" to have a *truth-maker* (Tiercelin 2011: 279). But how can one explain what that property consists in?

It is at this very point that Peirce's logical and metaphysical account of relations offers an original agenda for a convincing realistic and scientific metaphysics. It would be impossible, within the scope of this article, to present a detailed analysis of the type of dispositional realism Peirce endorses. Let me just note that it is close to the one I have tried myself to defend (Tiercelin 2011: 247-380), which relies, in a nutshell, on four main assumptions: 1) a basically causal theory of properties; 2) a conditional dispositionalist account of laws; 3) an emphasis not only on efficient causation but also on teleological causation; 4) a defense of some kind of *aliquidditism* (or *thin* essentialism) (Tiercelin 2011: 347 ff). The fourth assumption is of special interest here, in so far as the "relational" (rather than "substantival") and dispositional realism Peirce endorses allows him to avoid the "holistic" and "idealistic" consequences which threaten any kind of relationism, in which "substances" or "objects" always tend to disappear. Peirce saw the merits of "relational" over "substantival" realism, more in keeping with what contemporary science and logic tend to show, underlining the importance of relations and the limits met by a simple subject-attribute conception, as may be found in the old Aristotelian logic. But such a position, in his mind, was in no way opposed to, on the contrary it implied, some "thin" essentialism. In order to have a better grasp of this, it might be worth taking a quick look at what the scholastics, especially the Scotists, meant by essence, *quidditism* and *haecceitism* (Tiercelin 2011: 348 ff).

Indeed, Scotus did not defend *any* kind of essentialism. In particular, he followed Avicenna more than Aristotle in stressing the *neutrality* or irreducible and positive *indeterminacy* of the "Common Nature". For Avicenna, essence, as such, can indeed be viewed under two headings, in things and in the intellect, but more importantly, in its pure essentiality, as being *neither* universal *nor* singular. The essence or "Common Nature" is neutral or indifferent to any further possible determinations. There are formal or metaphysical realities which are not to be

viewed as we call today "primitive thisness" (Adams 1979), precisely because they are, so to speak, awaiting further physical and logical determination. Thus, it is less crucial to think of essence independently of the properties which belong to it properly, that is, in distinguishing the essence from what makes it a particular *substance*, than to show how what is more what I have called myself an *aliquid* than a *quidditas* or a *substratum* without substance is necessary in order, then, to ground, on the logical level, logical universality, and, on the physical level, the quiddity of things. So, for both Avicenna and Scotus (and Peirce follows them here), to be a realist means neither to hypostasize platonic essences, nor to develop a form of essentialism simply devoid of the Aristotelian substantialist shape: it is first and foremost to admit, in distinguishing logical reality and real community, the irreducibility of a Common Nature which, in itself, is neither universal nor singular, although it is universal *in the mind*, and singular *in the things* outside the mind (Tiercelin 2011: 351).

Quidditism is not an attractive position to hold nowadays. For causal structuralists, in particular, quiddities are a 'will o' the wisp': or a way to say that I could have been a poached egg, no matter, so long as my haecceity was present (Hawthorne 2001). But the scholastics had a different approach. *Haecceitas* was introduced by Scotus to differentiate the singular from the universal, or the Common Nature *formally*. In order to be clear about the various categories that populate our world and establish the right alphabet of being, we should not confuse the logical, the physical and the metaphysical levels of our investigation (which may reveal more than one or two kinds of "essential properties"). In particular, as Peirce was later to argue, even if material essences are *dispositional*, it does not necessarily follow that *all* dispositional properties are essential. The fact that "X is hard" needs not be essential to X, even though hardness is a dispositional property causing X to behave in certain predictable ways.

Peirce made a "twist" to the Scotistic position: against the too static view of essence as defined by the Subtle Doctor, he argued that it is not the behaviour of a thing but rather its *habit of behavior* that constitutes the *intelligible* nature or real essence (2.664). Such a habit is a general disposition affecting the way that an object *would tend to behave* under certain *types* of circumstances. Both philosophers distinguished between the essence and the activities of a thing. However, Scotus and the medieval logicians were just able to deal with propositions that involved monadic predicates (like '—is hard'), not with those involving relational predicates (such as '—is a lover of—', or '—gave— to—') (3.464 ff; Raposa 1984: 151). Hence, they were only able to account for specific *classes* or *collections* of things, each class being comprised of all the subjects bearing a particular monadic predicate, and for the relation of *similarity* (that is, the sharing of a 'common nature') existing between the members of a given class. Peirce's aim with his logical analysis of relations was to go further and to analyze relationships other than that of *resemblance* of a certain object to the various members of its class. For he thought it much more important to make out the way in which laws govern the interactions between objects within a *meaningful* process. Now, the analysis of such a process or "system" involved the use of dyadic and triadic predicates: to claim that "X is hard" is to do more than ascribe a particular quality; rather, it is to assert that under certain specifiable conditions, X will or rather *would* tend to behave in a certain specifiable manner. Thus, "hardness" is to be regarded as a dispositional property, and a real "habit" or "law" must govern the behavior of those

objects within which it inheres. Any monadic predicate is in fact a sort of degenerate relative. So, if we want to make sense of a universe in which there are not mere simple qualities or pure possibilities (*Firstness*, in Peirce's jargon), or mere actualized possibilities in terms of individual events or mere existential reactions (*Secondness*), we have to proceed in that way. In a universe manifesting only *Firstness* and *Secondness*, namely devoid of generality and thus of intelligibility, it might be appropriate to speak of such non-relational monadic predicates. However, even when one is confronted with nothing more than the case of an individual object enduring through time, real continuity is involved, and the properties that inhere in such an object are themselves "general" (1.411 ff; 1.427). If the relationships between a thing and its properties can only be defined by a real habit, a "would-be" operating within the actual world of objects and events (Raposa 1984: 152), what is decisive, is not so much to specify the generality that characterizes a collection of objects having some quality in common (what Scotus does), but to account for the infinite number of real possibilities, i.e. the real and continuous relationships that exist between any two members of a class, between an object and its successive actualizations in time, between the interacting fragments of a system. 'X gives Y to Z' is general not simply because the relational predicate ('– –gives— to—') can be applied to many different sets of ordered triads, but rather because it ranges over the members of *any* given triad. Thus, the type of relationship Peirce is interested in is different from the 'sameness' that defines the medieval genera and species. More than classes of givers, gifts and recipients, what counts is the *system that encompasses* the giver, the gift, and the recipient, and the laws or habits of behavior that govern their interaction. In all types of relationships however, even in relationships of resemblance, a real continuity exists between *realia*, and predicates must be universalized or 'projected' in order to range over the infinite numbers of possibilities, actualized and unactualized, that make up the continuum (Raposa 1984: 153). What Peirce underlines is not only that there are real relations, but that *relations comprise the real natures of things*. Habits account for an object's essential intelligibility. They govern objects by relating certain types of behavior to specific kinds of circumstances. Hence, the essence of a thing is defined not by any particular relationship or activity within which the thing actually participates, but by a general habit or causal power that determines those relations and activities to which, given the appropriate conditions, that thing would be disposed. Such a habit is not simply essential to, but rather, must be of the essence of the thing, namely it must be predicated of the thing *per se primo modo* (2.361; Raposa 1984: 154; Tiercelin 2011: 295).

From this, another lesson may be drawn: if the essence of a thing is no collection of properties, but rather a special "habit of action", or a "bundle of habits" for a "law-cluster", we may well need more than mere *efficient* causation to explain the way it exerts its causal power as a whole, to view the thing in terms of a final cause specifying the general patterns of behavior it will tend to manifest and become (so that the causal function of the essences of things may be defined in terms of both formal and final causation). And we may also have to view the *binding* ("cement" or "glue") of all the objects itself in terms of some *final* (or *intentional*?) causation (Ellis 2001; Tiercelin 2011). At all events, this requires a careful elaboration and determination of the exact role played both by dispositions and by laws of nature in the intelligibility of nature. As I have argued elsewhere, both seem to be needed: dispositions find their intelligibility in the conditional necessity of laws; but laws can only be a true description of the world, provided they are

grounded in what things can do (in a dispositional and not merely possibilistic sense) (Tiercelin 2011: 344).

5. Concluding Remarks

From this brief presentation of Peirce's logical and metaphysical account of relations, I think we can already note an interesting point, namely that, for the logician of Milford, much more than: should we view relations basically as internal or as external? Or: is it so easy to draw the line between what we intend as a "relation of reason" and a "real relation"? The crucial issues to be dealt with seem to be: what is the best logic for a correct account of the reality of relations? Which does not merely mean: do we need other signs than symbols, namely indices, icons? But rather: should we favor graphical logic, or even build an indexical or iconic logic? And from the metaphysical perspective: how can we make sense of *foundationism*? More precisely: what is, indeed, the real *fundamentum relationis*? And, not so much: "should we opt for relationism or substantialism?" but: "how can we frame a satisfactory dispositional realism?" Whether or not Peirce's options are right depends of course on the stance one takes on the trend to pursue in logic and, in metaphysics, whether or not one is convinced (as indeed I am) by the virtues of dispositional realism. So, in a few concluding remarks, let me suggest a few merits of the latter position—which also seems implied by Peirce's views on relations—over, in particular, various kinds of structuralism.

Indeed, a detailed account of Peirce's dispositional realism would show how much it has in common with contemporary structuralism, whatever its variants might be (Tiercelin 2011: 368-374). However, it is likely that Peirce would also oppose *ontic structural realism*, which, strengthened by an underdetermination of individuality, seems to become today 'The Metaphysics' of fundamental physics (non relativistic quantum mechanics, quantum field theory, and general theory of relativity mainly). As critics have observed, when pushed too far, structuralism tends to be counter-productive: if there is nothing in the world but structure, to what will it be opposed? In general, when one resorted to the term of structure in science, and profitably so, it was because one meant it as an entity with blank places which objects could occupy. But if the latter must be "reconceptualized" or are meant to have a mere "heuristic" function (French 1999: 204) or even to disappear (French and Ladyman 2003: 37), what role can the structure itself still play (Psillos 2006, 2011; Chakravartty 2003; Tiercelin 2011: 371)? Even more problematic is the fact that if dispositional realists may be willing to assert the non supervenience of relations on the objects, namely that objects do not have any existence or identity independently of the relations they have with one another, they are not ready to accept the pure disappearance of the objects, which is advocated by some eliminativist ontic structural realists. If relations are merely primary in relation to objects which are literally constituted by them, or simple "nods" within structures in a relation of asymmetric dependence, then there are no objects any more, only relations or structures, namely relations without *relata*. If relational structures are ontologically more fundamental than individual objects, then all there *is*, is structure. Now, several *reasons* (and not only common sense) seem to militate in favor of maintaining the category of "object" in our ontology. A metaphysical one, first: without *relata*, relations have no reason of being; even if such *relata* have not necessarily any intrinsic identity. Secondly, an empirical reason: the physical characteristics on which one relies do not in the

least suggest to abandon such a commitment for objects in the fundamental physical world. Finally, a logical reason, which has to do, as Esfeld and Lam (2010) have mentioned, with quantifying over objects in standard first order logic and the apparently unavoidable use of set theoretic concepts in physical theories. If one tries to pull too far the very meaning of our primitive concepts of "real" and "object", we run the risk of rendering the world simply unintelligible (Heil 2003: 58-60).

As is usually claimed, *causal* structural realism is, in many respects, more convincing, in particular in its "moderate form": while giving ontological priority to relations, it does not deny that properties and objects are part of a fundamental ontology; however such properties need not be intrinsic, they may be relational or extrinsic. If there are physical relations between objects or *relata*, such objects have themselves relational properties. While the universal context of entanglement and non separability in quantum mechanics is fully admitted, a principle *of weak discernability* is also granted and viewed as a symmetric and irreflexive relation between two objects (hence there are two objects and not only one), which has some merits over mere ontic structuralism: like in dispositional realism, properties are well identified through their causal roles and the structures are defined as a network of causal relations among properties, hence by the causal powers which they confer to their possessors. Yet, it remains to be shown how it handles a problem which any kind of dispositional monism has to face, when forced to follow a *holistic* model. *Causal* structuralism is indeed a structuralism that rejects any form of quidditism, or the view according to which there would be "something" beyond the causal profile which, independently of it, insofar as it might exist, would make of that property what it is. But if no property can be identified unless all the others are, it looks as if none of them can be identified *simpliciter*. We hoped to understand the identity of properties while avoiding unknowable quiddities, we have merely moved the problem to another place. Since what we come to is a holistic network of relations among properties which seems even more mysterious and which is not more able to identify the properties. Quiddities have not disappeared: they have become a global "totusity" (Psillos 2011). As Chakravartty observes, "any case of warranted attribution of a causal property is facilitated by some properties which are being known independently of a knowledge of their other effects" (2007: 136). This seems plainly to grant that, in all cases, the conditions of individuation of the causal powers which are assured by the place they occupy in the global network can only be so provided that the identity of some properties or relations is fixed independently of the place they occupy in that network. And it also means that causation itself must be a relation identified independently of the role it plays in the causal network, even if it runs then the risk, from the point of view of structuralism, to turn this time into some kind of "hypostructuralism" (see Psillos 2011, Tiercelin 2011: 374).

Again, there is something preposterous to consider that "one can in principle discover what properties are through the effects they produce" (Esfeld 2009: 184), and that this applies to "all" the properties. First, because, to suppose that the real is knowable, at least in principle, does not imply that "everything" in the real is. As Peirce noted, there are "ultimate" facts which any one, be he a man of science or any man in the street should take account of (1.405), and in particular, such isolated facts as do not imply any explanation whatsoever (7.200; 7.194). Secondly, one should never underestimate the length, the complexity and even the

tricks of the various chains through which we come to discover the causal properties, some being too far from one another, some being hidden by the screen some may constitute. Besides, even granted that the total network of the causal profiles might be knowable, how could we ever know that it is indeed such and such properties that play such and such a role in the totality? Finally, what the limitations of causal structural realism show, is also, to what extent it is illusory and mistaken to think that one can do, in the end, as Peirce also clearly saw, without *aliquidditism* (Tiercelin 2011: 347 ff.), at least if one's aim is to provide genuine identity conditions, allowing, in particular to distinguish between the essential and the accidental parts of causal powers and to say what the *fundamentum* of things consist in. One cannot be satisfied with mere modal or conceptual distinctions, even in a Spinozist guise. For more than conceptualism is needed to be able to say what a thing consists in, what its real being is. Such a real being, its identity, is what makes the thing, the thing it is. Indeed, any radical anti-essentialism would take us to such a global anti-realism that it would surely be incoherent, as E.J. Lowe rightly pointed out (2007: 92). Without a minimal essentialism, or a "serious essentialism" (2007: 86), neither in the sense of an ersatz essentialism of possible worlds or of an essentialism of act and potency, but capable of specifying, for each object, the very being of the reality it signifies, which was Locke's (*Essay*, III, 3, § 15) as well as Aristotle's definition, it becomes very problematic, not even to know but merely to *understand* what is at the root of the intelligibility of things.

If many of these suggestions are already implied, as I think they are, in Peirce's account of relations, and more generally in the scholastic realism he defends, then they are still worth being carefully studied and discussed by any serious metaphysician today.

References

Adams, R.M. 1979, "Primitive Thisness and Primitive Identity", *Journal of Philosophy*, 76, 5-26.

Chakravartty, A. 2007, *A Metaphysics for Scientific Realism: Knowing the Unknowable*, Cambridge: Cambridge University Press.

Dipert, R. 1995, "Peirce's Underestimated Role in the History of Logic", in K. Ketner (ed.), *Peirce and Contemporary Thought*, New York: Fordham University Press.

Dipert, R. 2004, "C.S. Peirce's Deductive Logic: Its Development, Influence and Philosophical Significance", in C. Misak (ed.), *The Cambridge Companion to Peirce*, Cambridge: Cambridge University Press, 287-324.

Ellis, B. 2001, *Scientific Essentialism*, Cambridge: Cambridge University Press.

Esfeld, M. 2009, "The Modal Nature of Structures in Ontic Structural Realism", *International Studies in the Philosophy of Science*, 23, 179-94.

Esfeld, M. and Lam, V. 2010, "Ontic Structuralism as a Metaphysics of Objects", in A. Bokulich, P. Bokulich (eds.) *Scientific Structuralism*, Dordrecht: Springer.

French, S. 1999, "Models and Mathematics in Physics", in J. Butterfield and C. Pagonis (eds.), *From Physics to Philosophy*, Cambridge: Cambridge University Press.

French, S. and Ladyman, J. 2003, "Remodelling Structural Realism: Quantum Physics and the Metaphysics of Structure", *Synthese*, 136, 31-56.

Haack, S. 1992, "Extreme Scholastic Realism: Its Relevance to Philosophy of Science Today", *Transactions of the C. S. Peirce Society*, 28, 19-50.

Hawthorne, J. 2001, "Causal Structuralism", *Philosophical Perspectives*: Metaphysics, 15, 361-78; repr. in *Metaphysical Essays*, Oxford: Oxford University Press, 2006.

Heil, J. 2003*, From an Ontological Point of view*, Oxford: Clarendon Press.

Lowe, E.J. 2007, "La métaphysique comme science de l'essence", in E. Garcia et F. Nef (eds.), *Métaphysique contemporaine: propriétés, mondes possibles et personnes*, Paris: Vrin, 65-120.

Martin, R.M. 1980, *Peirce's Logic of Relations and Other Studies*, Dordrecht: Foris Publications.

Mellor, D.H. 1991, *Matters of Metaphysics*, Cambridge: Cambridge University Press.

Merrill, D.D. 1978, "De Morgan, Peirce and the Logic of Relations", *Transactions of the Charles S. Peirce Society*, 14 (4), 247-84.

Michael, E. 1974, "Peirce's Early Study of the Logic of Relations, 1865-67", *Transactions of Charles S. Peirce Society*, 10, 63-75.

Peirce, C.S. 1931-58, *The Collected Papers of C.S. Peirce*, C. Harsthorne, P. Weiss, and A. Burks (eds.), Cambridge, MA: Harvard University Press (8 vols.).

Psillos, S. 2011, *Knowing the Structure of the World*, Basingstoke: Macmillan, Palgrave.

Raposa, M.L. 1984, "Habits and Essences", *Transactions of the C.S. Peirce Society*, 20 (2), 147-67.

Thibaud, P. 1975, *De l'Algèbre aux Graphes, La logique de Charles Sanders Peirce*, Aix: Presses de l'Université de Provence.

Tiercelin, C. 1991, "Peirce's Semiotic Version of the Semantic Tradition in Formal Logic", in N. Cooper and P. Engel (eds.), *New Inquiries into Meaning and Truth*, Hertfordshire: Harvester Press, 187-213.

Tiercelin, C. 1993, *La Pensée-Signe : études sur Peirce*, Nîmes: éditions J. Chambon.

Tiercelin, C. 2011, *Le Ciment des choses*, Paris: éditions d'Ithaque.

Externalist Thought Experiments and Directions of Fit

Casey Woodling

Coastal Carolina University

Abstract

The classic thought experiments for Content Externalism have been motivated by consideration of intentional states with a mind-to-world direction of fit. In this paper, I argue that when these experiments are run on intentional states with a world-to-mind direction of fit, the thought experiments actually support Content Internalism. Because of this, I argue that the classic thought experiments alone cannot properly motivate Content Externalism. I do not show that Content Externalism is false in this paper, just that it cannot be motivated by the classic thought experiments alone. I discuss various externalist responses to the argument I raise and show that they all fail.

Keywords: directions of fit, content externalism, content internalism, thought experiment, Twin Earth, Burge

Content Externalism holds that the content of intentional states is not determined solely by the intrinsic or non-relational properties of the subjects who have those states.[1] In short, intentional content fails to be completely determined by the ways a subject is independent of the environment in which he or she is embedded.[2]

[1] Brown (2004) rightly notes that Content Externalism breaks down into different varieties. In this paper, I mainly focus on the two versions that are motivated by the classic thought experiments, though I do discuss other versions of Content Externalism in the paper, namely Fodor's version of Content Externalism related to his Conceptual Atomism. One of the main versions of Content Externalism I focus on is Natural Kind Externalism, which is the idea that intentional states about natural kinds involve natural kind concepts whose individuation depends on factors outside of the subjects who have or grasp these concepts. The other version that I focus on is Social Externalism, which is motivated by Tyler Burge's work. This view is more encompassing than Natural Kind Externalism and holds that the concepts that structure our thoughts depend on our relationship to others and the wider community in which we are embedded. Thus, our intentional contents do not depend on intrinsic properties alone.

[2] This debate has also been put in terms of wide and narrow content. Narrow content is content that is completely determined by a subject's intrinsic properties while wide content is content that is not completely determined by a subject's intrinsic properties but depends

Argumenta 2, 1 (2016): 139-155
ISSN 2465-2334

DOI 10.14275/2465-2334/20163.WOO

Some philosophers are convinced of the truth of this doctrine because of the classic externalist thought experiments, such as Tyler Burge's arthritis case and the Twin Earth thought experiment. Here is Paul Boghossian on the powerful role these classic thought experiments have had in motivating Content Externalism.

> [P]hilosophers who embrace externalism don't do so because they regard it as a self-evident truth. They embrace it, rather, because their intuitive responses to a certain kind of thought experiment—Putnamian Twin Earth fantasies—appear to leave them little choice.[3]

The classic thought experiments ask us to consider cases that involve beliefs, which have a mind-to-world direction of fit.[4] In this paper, I show that when we consider intentional states with a distinct direction of fit, such as desires, the thought experiments actually motivate Content Internalism and not Content Externalism.[5] Looking at intentional states with world-to-mind direction of fit, then,

at some level on the extrinsic or relational properties of the subject. There are dual-factor theories of content, which hold a role for both types of content. Ned Block (1986, 1987) has been very influential in this regard. It is thought that such approaches could do justice to both Content Externalism and Content Internalism by allowing for both sorts of content. Narrow content is needed for psychological explanation and to explain the rationality of subjects while wide content is needed to respect the role that the environment plays in determining content. I myself endorse the view that intentional content is narrow and semantic content is wide. That is, I endorse a view that combines Content Internalism and Semantic Externalism. However, there is no direct argument for this view in the paper, nor is the truth of the view presupposed anywhere.

[3] Boghossian 1997: 163.

[4] Though she did not use the expression 'direction of fit', the conceptual distinction is usually credited to Anscombe 1957. The distinction and terminology also appear in Austin 1953.

[5] The two types of direction of fit I shall consider are mind-to-world direction of fit and world-to-mind direction of fit. These categories are not exhaustive. As John Searle (2004) notes, some intentional states have a null direction of fit. For the purposes of this paper, it is important to understand that states with mind-to-world direction of fit earn their name because the aim of these mental state is for the mind to fit the world. Beliefs are intentional states that clearly have a mind-to-world direction of fit. States with a world-to-mind direction of fit earn their name because the aim of these mental states is for the world to fit the mind. Desires are clear examples of this type of direction of fit. For what it is worth, Searle 2004 offers a brief diagnosis of the classic thought experiments involving beliefs. To the Twin Earth argument, he argues in favor of an internalist intuition that the satisfaction conditions for intentional states are set from the subject's point of view. I think this approach is largely correct, though he suggests that the satisfaction conditions for intentional states about what is called 'water' is largely consistent across the respective populations, a point with which I am not in complete agreement. I would want to allow for more diversity in terms of the intentional content across individual subjects. His diagnosis of Burge's thought experiment is more or less that the relevant intentional content of the two individuals is the same. The only difference is that in one case the patient's use of 'arthritis' diverges from the community norms. Searle says this divergence is fine, but not enough of a reason for thinking that the respective intentional contents are distinct. I am largely in agreement with this diagnosis of the original thought experiment, though I think that in the end more needs to said to address Burge's Social Externalism properly as it motivated by other reasons, namely that Social Externalism is true if our thoughts connect to an objective reality.

shows that Content Externalism cannot be motivated by the classic thought experiments alone. Though there are many interesting issues in the vicinity, my focus in the paper is rather narrow: to show that the classic externalist thought experiments do not support Content Externalism on their own. This conclusion should be acceptable to both those who see thought experiments as viable tools in philosophical argumentation as well as those who are more skeptical. For the non-skeptic, I present evidence that the classic thought experiments support Content Internalism when run on states with a world-to-mind direction of fit, so the externalist intuition does not hold over various directions of fit. Those who are more skeptical of thought experiments in philosophy can read the conclusion as providing more grist for the mill; there is genuine debate about which intuitive response to the classic externalist thought experiments is best, so we must conclude that the thought experiments alone cannot motivate either Content Externalism or Content Internalism. Either way, the classic thought experiments alone do not motivate Content Externalism.[6]

1. The Classic Externalist Thought Experiments

Here is the textbook version, from *The Stanford Encyclopedia of Philosophy*, of how Twin Earth can be extended from Semantic Externalism (a view about linguistic meaning) to Content Externalism (a view about the intentional content of thoughts).

> Although this thought experiment was designed to establish semantic externalism, it can be extended to mental contents as well (see McGinn 1977). Thus, consider an individual on Earth who sincerely utters 'water quenches thirst' before 1750. Such an individual would be expressing his belief that water quenches thirst, a belief that is true if and only if H_2O quenches thirst. The externalist then asks us to consider a physically identical counterpart of this individual on Twin Earth. Being a resident on Twin Earth, this counterpart has only encountered twin-water, and has never encountered samples of water or heard about water from other people. According to the externalist, our intuition tells us that this individual on Twin Earth does not believe that water quenches thirst. When he utters 'water quenches

[6] Those who are looking for an argument for the falsity of Content Externalism will not find one in this paper. Some of the most common arguments of this form appeal to epistemic notions such as privileged access, self-knowledge and first-person authority, holding that Content Externalism is not compatible with agents having these epistemic properties in the correct ways. I also endorse a version of this argument. In a nutshell, the problem is that some forms of Content Externalism face a dilemma on the assumption that the concept associated with a word is what expresses the meaning of that word. The dilemma involves different types of concepts, mental particulars and abstracta. Either the content externalist says that the concepts that constitute intentional content are mental particulars or abstracta. If it is the first option, then the content externalist has to reject the distinction between communal and idiolectic meaning. If the latter option, then the content externalist has to reject the idea that subjects have privileged access to content, since no individual has privileged access to abstracta. A final option for the content externalist is to hold that concepts are neither particulars nor abstracta but reducible to abilities. This is not a live option, though, for Social Externalism (Burge's form of externalism) or Natural Kind Externalism (externalism motivated by Twin Earth cases) as these views are not consistent with this view of concepts. Not all forms of Content Externalism face the dilemma, but the two major views discussed in this paper do.

> thirst', he is instead expressing the belief that twin-water quenches thirst, a belief with different truth-conditions. In short, these two individuals have different beliefs despite being intrinsically identical (ignoring the fact that the human body is about 60% water). It follows that some beliefs do not supervene on intrinsic facts, and therefore that externalism is true.[7]

The twins share all the same intrinsic properties, though their respective intentional states differ in terms of their truth conditions. One twin's belief that water is vital to human life is true if and only if H_2O is vital to human life, and the other twin's belief is true if and only if XYZ is vital to human life. The difference in the extension of 'water' at each world makes for distinct truth conditions that in turn make the content of their beliefs different. This difference in intentional content cannot be captured by the intrinsic properties of the twins for they share all the same intrinsic properties, so intentional content is not determined merely by a subject's intrinsic properties. Content Internalism is false.

At a minimum, our reflection on the twins and the different substances at their respective worlds shows us that the nature of the environment has a decisive impact on determining intentional content according to the most popular interpretation of the thought experiment. The most basic Twin Earth intuition appears to be: the difference in chemical structure of the watery stuff in each environment makes for a difference in the content of the twins' beliefs about the watery stuff in each of their respective environments. In part, we come to this intuition by way of the set-up of the thought experiment. It is built into the thought experiment that intentional content is determined by conditions of satisfaction (which includes truth conditions). I take the assumption that intentional content is determined by satisfaction conditions to be a harmless assumption in the context of the thought experiment.[8] The conditions of satisfaction of intentional states are the state of affairs in the world that would satisfy the intentional state. In the case of belief, this is the state of affairs that would make some belief true. Because desires are not true or false, the state of affairs that satisfies some desire constitutes the desire's conditions of satisfaction.

Through reflecting on the classic Twin Earth thought experiment, we see that the truth conditions or satisfaction conditions of intentional states can be determined in a rather straightforward way: we note the terms in the intentional state reports, note their respective extensions, and then determine the satisfaction conditions of the intentional states of the respective reports. The intuition seems to be a powerful one.

Tyler Burge offers the other classic externalist thought experiment in which he describes two different patients who both assert, "I have arthritis in my thigh." The patients are embedded in environments where 'arthritis' has distinct linguistic or conventional meanings. In the first patient's environment, it means a rheumatoid ailment exclusively of the joints, and in the second patient's environment it means a rheumatoid ailment of either the joints or the muscles. Because of this difference in linguistic meaning, there is a difference in the truth conditions of the

[7] Lau and Deutsch 2014.

[8] I am not saying here that the content of *all* intentional states is fixed by conditions of satisfaction. See Crane 2013 for resistance to the idea that all intentional content is determined by conditions of satisfaction.

respective beliefs, and therefore a difference in their intentional content. The individuals share all the same intrinsic properties, so intentional content fails to supervene on intrinsic properties. The dominant intuition about this case is that Burge's externalist read of the thought experiment is indeed correct.[9]

There is a minority report. In the context of describing a version of the thought experiment where Twin-Oscar comes to Earth, Noam Chomsky registers the contrary intuition.[10]

> Turning to 'content of belief', if Twin-Oscar continues to ask for what comes from the faucet to quench his thirst, calling it 'water', has he changed his beliefs about water—irrationally, since he has no evidence for such a change? Or is he behaving rationally, keeping his original beliefs about water, which allow for the stuff on Earth to be water (in Twin-English) in the first place? If the latter, then beliefs about water are shared on Earth and Twin-Earth, just as on either planet, beliefs may differ about the very same substance.[11]

Chomsky's intuition is that the beliefs of an individual switched between Earth and Twin Earth would not change based on environmental changes—a thought shared by other internalists. Instead of attempting to sort out these conflicting intuitions, I shall simply grant the externalist intuition about intentional

[9] Burge's defense of Content Externalism is sophisticated and subtle. Burge has been clear that his various externalist thought experiments are deeper than the idea that the semantic content of the ascription of an intentional state always faithfully determines the intentional content of that state. See Burge 2003, 2006 for discussion of this point. He allows room for malapropisms and other instances where subjects misspeak, where, in other words, the standard semantic content of a term or phrase does not properly express the subject's intentional content. When Yogi Berra said, for example, "Texas has a lot of electrical votes," Burge's view is not that we attribute to him the concept ELECTRICAL as a component of his intentional content. In short, the semantic content of ascriptions does not always completely fix a state's intentional content according to Burge. Any plausible version of Content Externalism will have to accept this fact, and Burge's of course does.

[10] Some of Chomsky's philosophical and linguistic commitments, namely his focus on I-language as opposed to E-language, may indeed skew his intuitions here. This is likely the case, given that Chomsky (1986, 1995) holds that I-languages should be seen as the proper object of linguistic study and these objects can be studied without regard to the environment in which subjects are embedded. This point opens up a more general worry about thought experiments: that they merely reveal prior commitments of theorists and do not provide evidence for theoretical commitments. If this is so, then we have a general reason to doubt that the externalist thought experiments motivate Content Externalism, as they would not provide evidence *for* a view but evidence that a view is antecedently held. I do not want to endorse this general skepticism. I bring up the point merely to say that the externalist cannot reject Chomsky's intuitions on these grounds while maintaining the conclusion that Content Externalism is properly supported by the classic thought experiments. Related to this point, Machery (2012) points to data that suggests that one's field (and therefore theoretical focus) can bias one's intuitions about reference. It appears that philosophers of language, for example, have more Kripkean intuitions than do sociolinguists who typically have more descriptivist intuitions. Machery's favored explanation of this data is that one's theoretical commitments bias one's intuitions. I thank an anonymous referee for helping me bring out this important point.

[11] Chomsky 2000: 149. For what it is worth, Chomsky also expresses a general skepticism about the viability of using such thought experiments to tell us anything interesting about language and the mind.

states with a mind-to-world direction of fit for now and ask whether or not the externalist intuition can be sustained over intentional states with distinct directions of fit, for it must be so sustained if the thought experiments are to properly motivate Content Externalism. Let us turn now to intentional states with a world-to-mind direction of fit.

2. Testing the Classic Externalist Thought Experiments with Desire

In this section I show that the classic thought experiments elicit internalist intuitions when we run them on intentional states with a world-to-mind direction of fit. Consider a variant of the switching case.[12]

> I desire a drink of water. According to the content externalist, on Earth my desire to drink water is satisfied if and only if I drink H_2O, just as my belief that water is vital to human life is true if and only if H_2O is vital to human life. Now suppose that I am switched without my knowledge to Twin Earth, per the familiar slow-switching scenario. Here I add one detail to the standard switching scenario. I have a bottle filled with water (that is, H_2O) that gets transported with me when I am switched. Some time passes, the concept TWATER takes hold such that when I assert, 'I want a drink of water', my desire is now satisfied if and only if I drink XYZ according to a content externalist. In an attempt to satisfy my desire I drink from the bottle that travelled with me from Earth. As far as I am concerned, my desire is satisfied by the event of my drinking from this bottle. It certainly seems to me that my desire is satisfied.

It turns out that my desire is not satisfied by the event of my drinking from the water bottle according to the externalist, because the satisfaction conditions of my desire are not met, for my desire is satisfied if and only if I drink XYZ. Imagine my surprise when I learn that I must drink another glass of what appears to me to be identical to stuff that I just drank in order to satisfy my desire.

Based on the original version of the thought experiment (where we note that the differences in the extension of 'water' make for a difference in intentional states truly described by ascriptions with that term), the content externalist analysis is that my desire can be satisfied only by XYZ and nothing else. So, it would seem that this case shows that the externalist idea that the satisfaction conditions of intentional states are fixed by environmental factors is false, because if Content Externalism is true, then once the new Twin Earthian concept TWATER takes hold, all my intentional states about watery stuff will have the concept TWATER as a constituent. It seems clear, though, that my desire for what I call 'water' will be satisfied by H_2O. After all, if this stuff satisfied my desire for water on Earth, why would it not satisfy my desire for water on Twin Earth? But this is not something that the content externalist who wants to use the thought experiment to motivate Content Externalism can allow for. The case shows that the conditions of satisfaction according to Content Externalism are too restrictive. It is perfectly obvious that H_2O (and not just XYZ) will satisfy my desire for what I call 'water'.

[12] The switching scenario was first introduced in Burge 1988. Burge (1988) articulates the commonly held externalist idea that it would take a certain period of unspecified time for a switched individual's concepts to switch over and begin to fit the new environment.

The switching variants of Putnam's original Twin Earth case are not part of the original thought experiment. Externalists divide on how to best handle these cases, so it is important to discuss these divisions. There are, broadly speaking, two approaches to the slow switching cases, the Conceptual Addition Interpretation (hereafter CAI) and the Conceptual Replacement Interpretation (hereafter CRI).[13] CRI holds that after the switch, the switchee's concepts are completely replaced. In other words, if I am switched from Earth to Twin Earth, then my concept WATER, after the requisite period, gets replaced completely by the concept TWATER. On the other hand, CAI holds that both concepts (WATER and TWATER) can be retained; which concept is deployed will depend on the context. It may seem that CRI is the version of externalism that gets the wrong reading of my thought experiment and therefore the externalist can simply accept CAI to avoid the internalist intuition.

This is a plausible move, because I have so far been assuming that content externalists adopt CRI. At this point, we can see that Content Externalism supplemented with CRI cannot be used by someone who wishes to use the classic Twin Earth thought experiments to motivate Content Externalism. But what about Content Externalism supplemented with CAI? If I can acquire the concept TWATER while retaining the concept WATER during my strange trip, then the content externalist can say that my desire for water is satisfied by H_2O because my desire deploys my concept WATER because the intentional state is in some way tied to my home environment.

It turns out, though, that Content Externalism plus CAI cannot be used by someone who holds that Content Externalism is motivated by the classic Twin Earth thought experiment. To see this point, consider the standard telling of the Twin Earth story, which considers beliefs about watery stuff had by twin individuals. For this scenario to motivate Content Externalism it must be the case that the concept that each individual associates with 'water' not be disjunctive between XYZ and H_2O, because if it were, the intentional content would clearly not be distinct for each individual. And we need the intentional content of the beliefs to be distinct in order to conclude that intentional content fails to be determined by the intrinsic properties of the subjects. So if the content externalist says that the concept associated with 'water' is disjunctive, then he can say the right thing about my examples, but he cannot use Twin Earth to motivate Content Externalism, since such use of the thought experiment requires that the intentional content be distinct for the twin individuals who have the same intrinsic properties.

The point here is not that the externalist is blocked from ever saying that a single concept can pick out distinct things. The problem for the content externalist in taking this option is related to the use of the classic thought experiments in motivating Content Externalism. In those thought experiments, the externalist cannot say that the relevant concepts are disjunctive because then the thought experiments cannot be used to motivate Content Externalism. A content externalist who adopts this line must say, then, that the thought experiments by themselves do not properly motivate Content Externalism, but that the view can be motivated otherwise, but this is in line with the very conclusion I am arguing for. My point is that Content Externalism cannot be motivated by the classic thought experiments alone.

[13] I use terminology from Parent 2013 here.

To this the externalist who adopts CAI may push back by saying that in the original thought experiment, the question of switching is not raised, so the issue of which concept associated with 'water' gets tokened does not even arise. But this reply misses something crucial about CAI: that the method of determining subjects' intentional content is more difficult than might first appear and involves more context than is available in the original telling of the thought experiment. In the original telling of the thought experiment, we move from the linguistic meaning of the term 'water' directly to the truth conditions of the respective intentional states. There is no consideration of other factors. It is stipulated that the twins' histories are the same, but these are in no way examined to determine the nature of their concepts, and this is precisely what CAI requires, as it embodies a method of content attribution that is subtle, requiring consideration of relevant context. So, in the end, an externalist who adopts CAI ultimately must admit that the original version of the thought experiment cannot alone motivate Content Externalism.

At this point, it is natural for the content externalist to shift to Burge's famous thought experiment and give up on the original version of the Twin Earth one, the idea being that the motivation for Content Externalism may come by way of Burge's thought experiment and the problems just mention can be simply bypassed. To begin to address this move, let us begin by considering a variant of Burge's famous arthritis case.

> Suppose that Burge's patient (the patient who, according to Burge, falsely believes he has arthritis in his thigh) also desires relief from pain in his thigh that he reports as being arthritis. He asserts, 'I want the arthritis in my thigh to go away'. As in the original telling, the patient is embedded in an environment where 'arthritis' refers exclusively to ailments of the joints. Also following the original version, we imagine a second patient, one who is an internal duplicate of the first but embedded in an environment where 'arthritis' refers to ailments of the muscles and the joints. The individuals are the same internally. Like the first patient, the second patient asserts, 'I want the arthritis in my thigh to go away'.

What are the externalist intuitions about the conditions of satisfaction of the respective desires? Perhaps the most natural thought is that the desire of the first patient is satisfied just in case the arthritic pain in the patient's thigh stops, and the same could be said for the second patient if we use their words as a guide for what they desire. However, things are more complicated, as 'arthritis' has different meanings in the different environments. In the version with states with a mind-to-world direction of fit, it turns out that, regarding the respective beliefs about what is called 'arthritis', the first patient's belief is false while the second patient's belief is true. So, analogously—stating the satisfaction conditions from our language community—in the present version with states with a world-to-mind direction of fit, the first patient's desire for relief is satisfied if and only if the arthritis in his thigh ceases, while the second patient's desire is satisfied if and only if the tharthritis in his thigh ceases. To make the distinction here vivid, let us suppose that the respective arthritic pain ceases in both cases. On the content externalist analysis, the first patient's desire is not satisfied while the second patient's is. They are the same in terms of intrinsic properties, and assuming that conditions of satisfaction fix intentional content, the respective contents are distinct. Thus, Content Internalism is false.

Is it plausible to attribute to the first patient a desire with conditions of satisfaction that cannot be satisfied because it is built into the intersubjective concept ARTHRITIS that arthritis cannot occur in the muscles?[14] There may be reasonable grounds for attributing to subjects conditions of satisfaction that cannot in fact be satisfied. If someone has a desire to find a very specific idealized soul mate, then we are warranted in attributing to him or her conditions of satisfaction that cannot be satisfied. If a child desires that Santa bring him a new pony, then we are likewise warranted. However, in the case of the first patient, it seems that he has a desire that can be satisfied, so there is no good reason for attributing to him an intentional state with conditions of satisfaction that cannot be satisfied. The internalist intuition, then, is that his desire is satisfied just in case the pain in his thigh that feels to him like arthritis ceases. This can be satisfied or unsatisfied, depending on how the world turns out. In fact, the internalist idea is that the intentional state of the second patient has the very same conditions of satisfaction.[15] The internalist reading attributes conditions of satisfaction that respect the subjects' respective interests on the world and also capture the proper level of detail in their thoughts about their pain. What's crucial to them is not the correct medical classification of their pain, but the pain itself and their prior experiences with such pain. The main reason for preferring the internalist interpretation to the externalist one is that it attributes conditions of satisfaction for the patient's desire about his pain that can be either satisfied or not satisfied. In some cases, as I noted, it is right to attribute conditions of satisfaction that cannot be satisfied (when subjects have desires about fictional entities, for instance). Surely the patient's desire is not about a fictional entity. It is about a very real pain in his thigh. So, the proper reflective response to the above variant of Burge's original thought experiment is an internalist one. Therefore, the version of Burge's thought experiment that involves desires does not support Content Externalism.[16]

The dialectical context is this. The thought experiments most commonly used to motivate Content Externalism elicit internalist intuitions when we consider intentional states with a world-to-mind direction of fit. Of the Twin Earth

[14] When necessary I distinguish between two types of concepts, following Laurence and Margolis 2007, intersubjective concepts that are abstracta and subjective conceptions that are mental particulars.

[15] Note that these intuitions may be arrived at only after a period of reflection. Many of these thought experiments are so unusual and complex that it takes a period of reflection to form one's judgment about them.

[16] It appears that determining the conditions of satisfaction for intentional states with a world-to-mind direction of fit cannot be done without regard to how such states relate to other intentional states in the subject's mental economy. So, it may be that considering intentional states with this direction of fit has ramifications for issues related to Conceptual Holism, Molecularism, and Atomism. This paper is neutral on which of these views is correct. However, a very quick argument for the truth of Conceptual Molecularism (or perhaps something more radical) runs as follows. The conditions of satisfaction for my desire for eating biscuits, say, depend on my beliefs about biscuits. I may have an idiosyncratic idea of what counts as a biscuit for instance. So, if the content of desires is cashed out in terms of their conditions of satisfaction, and those conditions of satisfaction depend for their fixing on other intentional states, then it would seem that at least Conceptual Molecularism would be established. I sketch this argument not to endorse it, but to show that considering intentional states with a world-to-mind direction of fit may be important to do across a range of views in the philosophy of mind. I thank an anonymous reviewer for helping me see this connection.

case, either the content externalist adopts CRI and cannot adopt the plausible read of the thought experiment (that is, my desire is satisfied by the water in my bottle) or the externalist adopts CAI and cannot hold that the original thought experiments motivate Content Externalism, as CAI admits that we need more context than we are given in the original example to determine the intentional content of the respective twins. Of the Burge case, the content externalist must attribute to the first patient a desire with conditions of satisfaction that cannot be satisfied. Sometimes, as I note, this makes sense, such as in cases of desires related to fictional entities. However, in the revised Burge case, on reflection, it is most plausible to attribute conditions of satisfaction that can or cannot be satisfied. So, to be clear, the conclusion at this point is not that Content Externalism is false or that Content Internalism is true; it is that one must appeal to more than the classic externalist thought experiments to support Content Externalism.

3. Content Externalist Replies

The content externalist could always accept the conclusion and hold that more than appeal to the thought experiments is needed to properly support Content Externalism. For those content externalists who wish to resist this conciliatory line of thought, there are various strategies for dealing with the variants of the famous thought experiments above. I shall review five replies and argue that all of them fail.

3.1 Limiting the Scope of Content Externalism to Merely Intentional States with a Mind-to-World Direction of Fit

A content externalist could grant that Content Externalism is true of just states with a mind-to-world direction of fit and not of states with a world-to-mind direction of fit. This concessive move does not work, because it has serious problems when it comes to explaining simple bits of reasoning behind intentional action. On this move we would say that when I am on Twin Earth *unawares*, I desire a drink of water and I believe that twater will come out of the faucet in my kitchen, and this belief and desire pair causes me to go to the kitchen for a drink of water. However, saying this is problematic. What is it that I go into the kitchen for? It does not appear that the externalist has a uniform answer to this question. In short, I would not be able to reason using intentional states with different directions of fit—even though ascriptions of these states would involve the same term in the content-clauses—for the concept expressed (the concepts WATER and TWATER) would vary depending on the direction of fit of the intentional state.[17] I need to be reasoning with the same concept in each one of the contents for the contents to properly link up and properly cause my action.[18]

[17] The same problem would arise in the Burgean cases. If the subject desires to rid himself of the arthritis in his thigh and believes that the doctor can help him get rid of it, then for those two intentional states to cause him to make a doctor's appointment, the intentional content expressed by 'arthritis' must be the same for both states.

[18] This problem is similar to the one raised in Boghossian 1992. Boghossian argues that if Content Externalism is true, then subjects cannot detect the validity of their reasoning a priori. So, Content Externalism cannot be squared with the obvious truth that the validity of reasoning should be detectable a priori.

3.2 The Kripkean Strategy

Another externalist reply is to hold that the internalist intuitions elicited above rest on conflating, in each case, two distinct desires. In the water bottle case, the H_2O from the bottle satisfies the subject's desire to quench his thirst, but it does not satisfy his desire for water. In the arthritis case, the patient's desire for the pain in his thigh to cease is satisfied if it ends, but the pain ceasing does not satisfy his desire for the pain from the arthritis in his thigh to stop. I call this the Kripkean strategy because it is similar to the strategy he employs in explaining modal illusion. According to Kripke, one might take oneself, for example, to have imagined water being XYZ, but what they have really imagined is not water but a water-like substance being XYZ. I think that Kripke's strategy applied to modal illusion is implausible, just as I think that this reply is implausible. Let me explain.

So far I have merely granted the externalist intuitions about intentional states with a mind-to-world direction of fit. Responding to this reply helps us to see that these intuitions are mistaken as well. Consider the water bottle case. What is the concept most appropriate to attribute to me when I desire a drink of water on Twin Earth? The descriptive concept CLEAR, COLORLESS, NEARLY ODORLESS AND TASTELESS LIQUID or the natural kind concept WATER which is the concept CLEAR, COLORLESS, NEARLY ODORLESS AND TASTELESS LIQUID AROUND HERE THAT HAS CERTAIN MICRO-STRUCTURAL PROPERTIES THAT DETERMINE ITS ESSENCE. Note that the natural kind concept TWATER has a similar description but it is distinct in virtue of the different microstructural properties of twater. The externalist says that it is not the descriptive concept but the natural kind concept that structures my intentional states. Why must we hold such a wooden view of concept attribution, though? Perhaps on different occasions I have states with merely the descriptive concept and perhaps I have states with the natural kind concept. Surely which concept is deployed—the concept WATER, the concept TWATER, or the concept CLEAR, COLORLESS, NEARLY ODORLESS AND TASTELESS LIQUID—depends heavily on the context and exactly how I am thinking about the objects in my environment. But the content externalist who motivates Content Externalism by using the classic thought experiments and holds CRI leaves no room for this sort of context dependence. For those externalist who adopt CRI, in the Twin Earthian environment, the concept TWATER constitutes in part the subject's intentional content. Period, end of story. The point of my variation of the thought experiment is to bring out that sometimes the subject's interest in the world means that the descriptive concept and not the natural kind concept gets tokened. If I want to drink water, it typically does not matter to me whether it is H_2O or XYZ given that these microstructures give rise to the very same functional and appearance properties. Upon drinking the water in my bottle, I take my desire to be satisfied and it is. Upon being told the whole story about water and twater, I will not revise my belief that my desire is satisfied. The externalist read is that I will admit that I was wrong about whether my desire was satisfied, but I do not think that this is right. It seems open for me to retain my belief that my desire was satisfied given that I got the very thing that I was looking for. In this case, my thinking about the liquid in my environment was not as detailed as the natural kind concept WATER. Surely, we should attribute to subjects concepts that most closely capture how they think of the world. Content Externalism plus CRI does not allow for this, because to all subjects who have intentional states described by

ascriptions involving 'water' on Twin Earth, we must attribute the concept TWATER (after of course the requisite period of time passes).

Again, those content externalists who endorse CAI have resources to allow for more sophisticated interpretations of intentional content; however, the point for our purposes is that these individuals—due to these very resources—cannot use only the classic thought experiments to motivate their view. They must allow that we do not have enough context in the original thought experiment to determine the nature of the respective intentional contents.

In the Burge case, which concept structures the patient's thought? The concept ARTHRITIS or the concept PAIN THAT FEELS LIKE ARTHRITIS FEELS? Burge's view is that it is the former, in part because the patient uses 'arthritis' to describe his own intentional state, and this expresses the intersubjective concept ARTHRITIS in this situation. Running the case with desires helps us to better see that it is much more natural and intuitive to think that the concept PAIN THAT FEELS LIKE ARTHRITIS FEELS structures his thought because it allows that his desire is satisfied when the pain goes away. What is crucial to his perspective is that his pain, which feels to him like arthritis, ceases. So, we need not attribute to him the concept ARTHRITIS in this case because he is clearly not thinking of his pain according to the intersubjective concept ARTHRITIS, as this would require that he thinks he has a pain in his thigh that cannot occur in his thigh. As with the case of WATER and TWATER, Content Externalism can be seen to have a much too restrictive view of the concepts that structure the thoughts of subjects. Sometimes we think with natural kind concepts but sometimes we do not. And what makes the differences is not any environmental factors but factors about our interest in the world. Running the thought experiments on desire gives us a new perspective on these thought experiments and helps us to see that Content Externalism supplement with CRI is too rigid and restrictive when it comes to the sort of concepts that we can attribute to subjects.

3.3 Conceptual Holism, Molecularism, and Atomism

It may seem that the content externalist has other resources to use in responding to the thought experiments involving states with a world-to-mind direction of fit. One thought is that these experiments assume a version of Conceptual Holism or Molecularism and thus do not trouble the conceptual atomist.[19] The defining feature of Conceptual Atomism is that concepts have no internal structure. They

[19] Conceptual Holism is the analog of Meaning Holism. On such a view, the content of any concept depends on its relation to potentially all concepts in an agent's mental economy. Conceptual Molecularism is a restricted form of Holism. A concept has its content in virtue of the relations it stands in to a restricted range of concepts. Unlike these views, which share the idea that conceptual content is interdependent between concepts, Conceptual Atomism holds that concepts have content solely in virtue of relations they bear to objects in the environment. Fodor is the most prominent defender of this view (see Fodor 1987, 1990 and 1998 for classic discussions and defenses). It would seem that Conceptual Holism and Conceptual Molecularism are perhaps more suited to Content Internalism while Conceptual Atomism is more suited to Content Externalism, given that the former focus primarily (though perhaps not exclusively) on internal relations for fixing intentional content while the latter focuses exclusively on external relations for fixing intentional content. I am not saying that some of these views are incompatible with others, but merely suggesting that Holism and Molecularism seem prima facie more suited to Content Internalism while Atomism seems prima facie more suited to Content Externalism.

have no components, and get their content from relations that they bear to the environment. Fodor's Asymmetry Dependency Theory has it that concepts have the content they do in virtue of the external factors that typically cause them to be tokened. Of course, this view has historically struggled with explaining misrepresentation. For example, I may think I see a cow in the distance, and deploy the concept COW when there is in fact not a cow in the distance but a dog. To this Fodor holds that the concept COW means cow and not dog because the concept would not be tokened by a person who sees a dog unless it were usually tokened when a person sees cows. Unless my concept COW had been caused by regular interaction with large domesticated animals, then it would not be able to be mistakenly deployed (as it may be on rare occasion) when I see a dog.[20] Obviously, the question of which one of the above views of concepts is correct is outside of the scope of this paper. It is also worth noting that while Conceptual Atomism is taken to be a type of Content Externalism, it is not a form of Content Externalism that is motivated by the classic thought experiments.[21] These thought experiments motivate Natural Kind Externalism and Social Externalism, so it is worth noting that the conceptual atomist-cum-content externalist is not really troubled by anything I say in this paper. My target are those individuals who hold that Natural Kind Externalism or Social Externalism can be motivated by the classic thought experiments alone.

It is worth bringing out, though, that none of my discussion above assumes that concepts are structured entities. I happen to think that they are, but nothing hangs on that above. I do discuss descriptive concepts and thereby discuss concepts that the conceptual atomist would clearly reject. However, I do not beg the question against the externalist by assuming that all concepts are descriptive concepts; I merely examine them alongside other types of concepts and ask the reader to reflect on their plausible deployment in certain situations.

3.4 The New Thought Experiments Are Too Complicated

Another reply is that the versions of the classic thought experiments that I have offered are so strange and complex that they fail to elicit any solid intuitions. Due to their complexities there simply are no natural or intuitive responses. The content externalist who raises this worry would have to bring the point home and say that the original versions of the thought experiments, as well as the switching variants, are too strange and complex to motivate Content Externalism in the first place. These are all very strange situations that admittedly stretch the normal application conditions of our concepts. So, this move by the content externalist, in the end, admits that I am right: something other than the classic thought experiments must be used to properly support Content Externalism.

3.5 Additional Worries, Additional Distinctions

Let me address a final potential worry. One might worry that I am conflating conditions of satisfaction being satisfied based on the subject's interests and the subject believing that conditions of satisfaction are satisfied. There are no doubt cases where subjects believe that their desires are satisfied while they are not. It is crucial to distinguish between an intentional state being "satisfied" because its

[20] See Fodor 1987, 1990.
[21] See Rives 2010.

subject believes it to be and the conditions of satisfaction actually being satisfied. Consider an example where these notions come apart. Suppose I set out to buy my wife a diamond necklace. I go to the jeweler and get a great price on what I believe to be a diamond necklace. Of course, my desire to buy my wife a diamond necklace is satisfied if and only if I buy her a diamond necklace. In this case it is very important to me that the necklace have diamonds and not a superficially similar material. The jeweler it turns out has tricked me: there is not a single diamond in the necklace—just cubic zirconia. So, I believe my desire is satisfied, but it is really not. Perhaps the above water-bottle case is like this, the externalist may argue: I believe my desire is satisfied, but it is really not. Surely my taking my desire to be satisfied is not enough for it to be truly satisfied. The variants of the classic thought experiments I offer do not assume that whether desires are satisfied depends on whether subjects believe that they are. Here is how to tell the difference between the conditions of satisfaction being fixed from the subject's perspective and actually being satisfied and being "satisfied" merely because the subject believes they are. In the case where I buy a cubic zirconia, once it comes to my attention that the necklace contains no diamonds, I will correct my belief that my desire has been satisfied. From my point of view, the desire for a diamond necklace is satisfied if and only if the necklace I buy contains real diamonds. (Unlike the water bottle case, the microstructure here matters very much and is relevant to my interests.) It seems that I will make no such correction when it to come to my attention that the liquid I drank was H_2O and not XYZ. It seems that I will maintain that my desire for water was satisfied by the H_2O in my bottle—even when I am given the full information about the distinct chemical composition of the types of watery stuff in my environment. From my point of view, the desire for what I call 'water' will surely be satisfied by H_2O. Whether my desire for water is satisfied is not determined by whether I believe it to be satisfied. The fact that I can meaningfully examine whether my belief about the satisfaction of my desire is true or false when I become aware of the chemical make-up of the liquid in my bottle shows that there are facts of the matter outside of my believing some way or other which settle the matter about whether the conditions of satisfaction are met. Rather than my desire being satisfied by my belief that it is, it is satisfied because the conditions of satisfaction are met. Even though the conditions of satisfaction are grounded in the subject's perspective, it should be clear that this is not the same thing as saying that the desire is satisfied just in case the subject believes that it is.

The same test can be applied to the arthritis case. Provide the subject in question with the full information about the situation, and then ask whether he would change his belief about whether his desire was satisfied. Suppose that the pain in the subject's thigh ceases; from his perspective, then, his desire becomes satisfied. Now suppose that he also becomes informed that 'arthritis' in his community means a rheumatoid ailment exclusively of the joints. Does he then change his mind about whether or not his desire was satisfied? Surely he will continue to believe that it is satisfied even when informed about his misuse of the term 'arthritis'. He will say that he mistakenly used the term 'arthritis' to describe his desire, but he will not revise his belief that his desire for relief from the arthritis-like pain in his thigh is satisfied. His pain, after all, is gone. His merely believing that his desire is satisfied does not make it satisfied. As in the previous case, the

desire is satisfied because the conditions of satisfaction are fixed from his perspective on the world and these conditions of satisfaction are met when the pain in his thigh—that feels to him like arthritis—ceases.

Though it is not novel, consider a final point to make the project of this paper more acceptable to philosophers who consider themselves externalists. As for the first water-bottle case I discuss, although my desire is not satisfied by XYZ as the content externalist says it is, it can still be true that the meaning of 'water' at Twin Earth is distinct from the meaning of 'water' at Earth. It may be that the semantic content or linguistic meaning of the intentional state ascription, 'I desire a drink of water', differs from place to place. If I am switched unawares between Earth and Twin Earth, I may mean one thing (in the sense of linguistic meaning) when I report my desire on Twin Earth and mean another when I report my desire on Earth, while all the while the intentional content of my desire remains the same in both places. A view such as this would be a combination of Semantic Externalism and Content Internalism and allow that the semantic content of some intentional state ascriptions can fail to properly describe the intentional content of such states. In the first Twin Earth variant, such a view says that I misdescribe my desire as one for water, because 'water' as uttered on Twin Earth is understood by the semantic externalist to refer exclusively to XYZ and my desire is clearly satisfied by H_2O. So, the semantic content of the language of my report is that I desire XYZ. However, this semantic content fails to describe the intentional content of the desire because the desire is clearly satisfied by H_2O and not merely XYZ.[22]

4. Conclusion

Although the classic externalist thought experiments typically appeal to just beliefs, we should test our intuitions about the classic externalist thought experiments on other intentional states with distinct directions of fit. Although I have focused on desires here, it seems that other types of intentional states with a world-to-mind fit would cause us to have internalist intuitions as well. Consider a modification of the first water-bottle case above. Suppose I hope that there is some water in my backpack. My hope is satisfied just in case there is some water in my backpack. Does the H_2O in my backpack satisfy my hope? It seems obvious that it does. The content externalist who wants to use the classic thought experiments alone to motivate Content Externalism says that the hope is satisfied only by XYZ if I have been switched to Twin Earth and been embedded in that environment long enough.[23] This is surely the counterintuitive verdict. Our intuitions tell us that the hope for what I call 'water' can indeed be satisfied by H_2O.

We can draw a moral at this point: things go wrong when we fix the satisfaction conditions of intentional states without proper consideration of the subject's perspective and interests about the world in some context, and we can see this point more readily when we reflect on intentional states with a world-to-mind

[22] See Loar 1988, Ludwig 1996 and Bach 1997 for similar ideas about how semantic content can at times fail to properly capture a subject's intentional content. My points here should not be understood as being novel, as the distinction between semantic and intentional content has been in the literature for some time. I merely use the distinction to bring clarity to the discussion, as sometimes it is not always kept in mind.

[23] Remember that only the content externalist who endorses CRI can use the classic thought experiments to motivate Content Externalism.

direction of fit. Our focus when we reflect on states with a world-to-mind direction of fit is first on how the subject conceives things as opposed to when we reflect on states with a mind-to-world direction of fit where there is a greater temptation to focus first on how the world is and only secondarily on how the subject conceives of the world or takes it to be.

Running the thought experiments on desires and other states with a world-to-mind direction of fit helps us see that Content Externalism cannot be properly motivated by the thought experiments alone. Let me end by saying that nothing I have said impacts the Twin Earth thought experiment's ability to support Semantic Externalism—the view it was originally designed to support. Perhaps the reported strength of the externalist intuition so often discussed in the literature is a result of running together the thought experiments' ability to support Semantic Externalism with their ability to support Content Externalism. We should be mindful of this, of course, and run the thought experiment separately for each version of externalism and also run it on a variety of intentional states and not just beliefs.[24]

References

Anscombe, E. 1957, *Intention*, Cambridge, MA: Harvard University Press.

Austin, J. 1953, "How to Talk—Some Simple Ways", *Proceedings of the Aristotelian Society*, 53, 227-46.

Bach, K. 1997, "Do Belief Reports Report Beliefs?", *Pacific Philosophical Quarterly*, 78, 215-41.

Block, N. 1986, "Advertisement for a Semantics for Psychology", *Midwest Studies in Philosophy*, 10, 615-78.

Block, N. 1987, "Functional role and truth conditions", *Proceedings of the Aristotelian Society*, 61, 157-81.

Boghossian, P. 1992, "Externalism and Inference", *Philosophical Issues,* 2, 11-28.

Boghossian, P. 1997, "What the Externalist Can Know A Priori," *Proceedings of the Aristotelian Society*, 97, 161-75.

Brown, J. 2004, *Anti-individualism and Knowledge*, Cambridge, MA: MIT Press.

Burge, T. 1979, "Individualism and the Mental", *Midwest Studies in Philosophy*, 4, 73-122.

Burge, T. 1988, "Individualism and Self-Knowledge", *Journal of Philosophy*, 85, 649-63.

Burge, T. 2003, "Phenomenality and Reference: Reply to Loar", in Hahn, M. and Ramberg, B. (eds.), *Reflections and Replies*, Cambridge, MA: MIT Press, 435-50.

Burge, T. 2006, "Postscript to 'Individualism and the Mental'", in T. Burge, *Foundations of Mind*. Oxford: Oxford UP, 151-81.

Chomsky, N. 1986, *Knowledge of Language: Its Nature, Origin and Use*, Westport, CT: Praeger.

[24] My thanks to two anonymous referees for insightful comments. Also, my thanks to participants at a session of the 2014 Meeting of the Southern Society for Philosophy and Psychology where I presented some of the ideas in this paper.

Chomsky, N. 1995, "Language and Nature", *Mind*, 104, 1-61.

Chomsky, N. 2000, *New Horizons in the Study of Language and Mind*, New York: Cambridge University Press.

Crane, T. 2013, *The Objects of Thought*, New York: Oxford University Press.

Fodor, J. 1987, *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, Cambridge, MA: MIT Press.

Fodor, J. 1990, *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press.

Fodor, J. 1998, *Concepts: Where Cognitive Science Went Wrong*, New York: Oxford University Press.

Lau, J. and Deutsch, M. 2014, "Externalism About Mental Content", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Zalta, E. (ed.), URL = http://plato.stanford.edu/archives/sum2014/entries/content-externalism/

Laurence, S. and Margolis, E. 2007, "The Ontology of Concepts: Abstract Objects or Mental Representations?", *Noûs*, 41, 561-93.

Loar, B. 1988, "Social Content and Psychological Content", in Grimm, R. and Merrill, D. (eds.), *Contents of Thought*, University of Arizona Press, 99-110.

Ludwig, K. 1996, "Singular Thought and the Cartesian Theory of Mind", *Noûs*, 30, 434-60.

Machery, E. 2012, "Expertise and Intuitions about Reference", *Theoria*, 37, 37-54.

McGinn, C. 1977, "Charity, Interpretation, and Belief", *Journal of Philosophy*, 74, 521-35.

Parent, T. 2013, "Externalism and Self-Knowledge", *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Zalta, E. (ed.), URL = http://plato.stanford.edu/archives/sum2013/entries/self-knowledge-externalism/

Rives, B. 2010, "Jerry Fodor", *Internet Encyclopedia of Philosophy*, URL = http://www.iep. utm.edu/fodor/

Searle, J. 2004, *Mind: A Brief Introduction*, New York: Oxford University Press.

Putnam on Methods of Inquiry

Gary Ebbs

Indiana University, Bloomington

Abstract

Hilary Putnam's paradigm-changing clarifications of our methods of inquiry in science and everyday life are central to his philosophy. He takes for granted that the judgments of scientists are for the most part reasonable and not in need of philosophical support, and that no part of our supposed knowledge is unrevisable or guaranteed to be true. He infers from key episodes in the history of science that our language contains terms whose references may remain unchanged despite radical changes in our theories, and that some statements are so basic for us at a given time that it would be unreasonable to give them up at that time, even if our failure to be able to conceive of alternatives to them is no guarantee that they are true. These central methodological commitments lead him to theorize that meanings are not in the head, that there are empirically discoverable property identities, and that reference is the key to understanding truth and realism.

Keywords: apriority, Carnap, functionalism, inquiry, meaning, Quine, realism, reference, Reichenbach, science, trans-theoretical terms, truth

Hilary Putnam died on March 13, 2016, at the age of 89. At the heart of his vast philosophical legacy lie his fresh, brilliant, and paradigm-changing clarifications of our methods of inquiry in science and everyday life.

Putnam's career began in the 1950s, an exciting time for philosophy in the United States. A series of revolutionary breakthroughs in logic, mathematics, and physics had recently prompted a new generation of thinkers to reconceive the relationship between philosophy and the sciences. These new thinkers, among them Hans Reichenbach, W.V. Quine, and Rudolf Carnap, Putnam's main mentors in graduate school and the early part of his career, announced that, contrary to what many philosophers, among them Descartes, have claimed, the judgments of scientists are for the most part reasonable and not in need of philosophical support.

This bold new attitude toward science is integral to Putnam's philosophy. He argues, for example, that the supposed paradoxes of time-travel dissolve when we employ the techniques of physics, and that the commonsense view that future events are as yet undetermined, hence less real than present events, is refuted by special relativity. In the latter case, Putnam concludes, "the problem of the reality and determinateness of future events is now solved.… [and] it is

Argumenta 2, 1 (2016): 157-161
ISSN 2465-2334

DOI 10.14275/2465-2334/20163.EBB

solved by physics, not philosophy" (Putnam 1967: 204). He digs into the details of contemporary theories of space-time to challenge Reichenbach's claim that these theories rest on conventional definitions, such as definitions of straight lines as light ray paths and congruence in terms of transport of rigid rods. For many years Putnam also argued that one lesson of quantum mechanics is that we need to give up the distributive laws of truth-functional logic. Though he later changed his mind on this point, he always agreed with Quine that no part of our supposed knowledge, no matter how clear it seems to us or how firmly we now hold it, is unrevisable or guaranteed to be true; and that insofar as traditional philosophical conceptions of reason, justification, and apriority conflict with this principle, they should be abandoned.

In one of his most important early papers, "The Analytic and the Synthetic" (Putnam 1962a), Putnam criticizes Carnap's analytic-synthetic distinction in ways that both challenge and extend Quine's earlier arguments against it. Before the development of relativity theory, Putnam explains, physicists were unable to see any way in which '$e = \frac{1}{2} mv^2$', an equation for kinetic energy, could be false. They held it immune from disconfirmation by new empirical evidence, and it was reasonable for them to do so. By Carnap's logical empiricist principles, Putnam notes, the methodological role of the equation is best explained by describing it as true by definition of kinetic energy. After Einstein developed relativity theory, however, scientists revised '$e = \frac{1}{2} mv^2$', replacing it with a more complicated equation that fits the new theory, and concluded that '$e = \frac{1}{2} mv^2$', while approximately true, is strictly speaking false, hence not true by definition.

To make sense of such cases, Putnam introduces the idea of a "law-cluster" term, which figures in many different laws of a theory. He observes that we can give up some of the laws in which such a term figures without concluding that the reference of the term has changed. For instance, we can continue to use a given term to refer to kinetic energy while radically changing our theory of kinetic energy. Such terms are, in a word, *trans-theoretical*.

In another of his ground-breaking early papers, "It Ain't Necessarily So" (Putnam 1962b), Putnam presents an example that challenges not only Carnap's logical empiricist principles, but also a wide range of more traditional conceptions of the role of reason in inquiry. Putnam observes, for instance, that our theories of the geometry of physical space have changed since the eighteenth century, when the principles of Euclidean geometry were so fundamental to our way of thinking about physical space that we could not then conceive of any alternatives to those principles. This may seem at first to suggest that when we developed alternatives to Euclidean geometry for physical space, we also thereby changed the meanings of the terms that we used to describe physical space, in a sense of "change the meanings" that implies that it would be incorrect to regard those terms as trans-theoretical, as retaining their reference despite the radical changes in our theory of physical space. Putnam rejects this response. In a characteristic passage that demonstrates his disarmingly direct and clear way of thinking about difficult technical topics, he writes,

> [Modern physics says that] our space has variable curvature. This means that if two light rays stay a constant distance apart for a long time and then come closer together after passing the sun, we do not say that these two light rays are following curved paths through space, but we say rather that they follow straight paths and that two straight paths may have a constant distance from each other for a

> long time and then later have a decreasing distance from each other. […] If anyone wishes to say, "Well, those paths aren't straight in the old sense of 'straight'," then I invite him to tell me *which* paths in the space near the sun are "really straight." And I guarantee that, first, no matter which paths he chooses as the straight ones…[they] will look crooked, act crooked, and feel crooked. Moreover, if anyone does say that certain non-geodesics are really straight paths in the space near the sun, then his decision will have to be a quite arbitrary one; and the theory that more or less *arbitrarily* selected curves near the sun are "really straight" […] would certainly *not* be a mere decision to "keep the meaning of words unchanged" (Putnam 1962b: 242).

Putnam argues that while our theory of physical space has changed radically since the eighteenth century, it is nevertheless correct to regard the terms that scientists in the eighteenth century used to refer to paths through physical space as trans-theoretical and to conclude that many of the sentences about physical space that scientists accepted in the eighteenth century, such as "Physical space is Euclidean," are false.

Putnam thinks there is an important *methodological* lesson to be learned from this case: some statements are so basic for us at a given time that it would not be reasonable to give them up at that time, even if our failure to be able to conceive of alternatives to them is no guarantee that they are true. As I mentioned earlier, Putnam thinks scientific judgments, even those to which we see no coherent alternatives, need no special philosophical justification. He concludes that if a person cannot specify any way in which a statement *S* may be false, it is reasonable for her to accept *S* and hold it immune from disconfirmation. An immediate consequence of this conclusion is that

> The difference between statements that can be overthrown by merely conceiving of suitable experiments and statements that can be overthrown only by conceiving of whole new theoretical structures—sometimes structures, like Relativity and Quantum Mechanics, that change our whole way of reasoning about nature—is of logical and methodological significance, and not just of psychological interest (Putnam 1962b: 249).

Putnam returns to this central methodological point repeatedly, exploring and clarifying it from many different points of view.

Putnam's compelling observations about theory change, first published in the 1960s, discredited the then standard theories of reference and meaning. His proposal that we view some of our terms as law-cluster (i.e. trans-theoretical) terms was a first step away from standard theories. A second step was to extend his notion of trans-theoretical terms, which he first introduced primarily to make sense of cases in which a single inquirer *changes* her view from one time to another, to cases in which two or more inquirers (or speakers) *simultaneously* use a term with the same reference despite large differences in the theories or beliefs they associate with the term. In this key step Putnam observes that we typically assume that ordinary English speakers can use the term 'elm' to refer to elm trees, and 'beech' to refer to beech trees, even if they know very little about elms and beeches, and cannot tell them apart. To distinguish elms from beeches, or to learn about these trees, such ordinary speakers rely on others who know more about them. We rely, in short, on what Putnam calls the division of linguistic labor. He proposes that we reject any theory of reference that implies that ordi-

nary speakers cannot refer to (or think about) elms when they use the term 'elm', even though they do not know much, if anything, about elms, except, perhaps, that they are trees. He also argues that the references of such "natural kind" terms are dependent, in part, on the environment in which they are applied, even if it takes years of inquiry and theorizing to discover what the references are. Finally, he argues that to discover the reference of a term and learn about its properties is also to clarify what it is true of, and thereby also to clarify one key component of the meaning of the term, namely, its contribution to the truth conditions of sentences in which it occurs. He theorizes that the meanings and references of a speaker's words are determined in part by causal relations the speaker bears to other speakers in her community and to the environment in which she applies the terms. All these points lead him to his famous conclusion that "meanings ain't in the head" (Putnam 1975a: 227).

Putnam's paradigm-changing views of meaning and reference discredit previously standard views of properties, according to which two terms with which a speaker associates different criteria of application must express different properties. Putnam's view of meaning and reference instructs us to focus not on the criteria of application that speakers associate with those terms—criteria that may vary from speaker to speaker even for the same term, and that, in any case, according to Putnam, do not determine reference—but on the things to which the terms are actually applied. With this shift in focus, it became possible to see how there might be *empirically discoverable property identities*, such as the identity of the property of being a portion of pure *water* with the property of being a suitably large clump of contiguous H_2O molecules.

This new view of properties smoothed the way for Putnam's enormously influential hypothesis that certain types of mental properties, such as the properties of desiring food, of believing that food can be found in the next room, or even of being in pain, are identical with Turing-machine computational-functional properties. Unfortunately, as Putnam himself later pointed out, since meanings aren't in the head, his Turing-machine functionalism fails to capture such ordinary mental properties as desiring a drink of water, or believing that elm trees are deciduous. He eventually concluded that most of the explanations of behavior that matter to us in everyday life "[cannot] be reduced to any of the various levels of description of the functioning of our neurons, including the computational level" (Putnam 2015: 59). Putnam continued to believe that there is something right about the idea that to be in a mental state is to be in a functional state, but he opted for a "liberal functionalism" that makes essential use of commonsense and scientific vocabulary, such as "desires a drink of water," and "believes that water can be found in the next room," to ascribe "capacities to function" that "reach out to the environment" (Putnam 2012: 83).

Throughout his career (with occasional lapses that he later regretted—see Putnam 2015: 90-92) Putnam was committed to scientific realism and to realism about inquiry more generally. He took reference to be key to understanding truth, and hence key to understanding realism. He rejected efforts by Quine and others to deflate questions about reference and truth by replacing these notions with surrogates defined using techniques from mathematical logic. The problem that Putnam returned to again and again, in different forms, is not that the replacements fail to capture the supposed concepts of truth and reference, but that they fail to incorporate trans-theoretical terms, which Putnam sees as integral to our methods of inquiry. According to Putnam, no theory of truth or reference

that satisfies certain basic constraints, including the constraint that it incorporate trans-theoretical terms, can do without any appeal to norms of truth and meaning. For this reason, among others, he concludes that normativity cannot be purged from our understanding of truth, reference, or meaning, or, ultimately, from our understanding of inquiry itself.

This last point might seem to give aid and comfort to those who think philosophy is an *a priori* discipline, whose results are higher or firmer than anything one can learn from science. Putnam has no sympathy for this kind of philosophical recidivism. It ignores one of the key lessons of his investigations of radical theory changes in science: no part of our supposed knowledge, no matter how clear it seems to us or how firmly we now hold it, is unrevisable or guaranteed to be true. The way forward, Putnam thinks, is not to revive a belief in a special source of a priori knowledge, but to engage, instead, in serious and honest inquiries into methodological roles of statements in all the disciplines and practices that weigh with us, including not only mathematics and the natural sciences, but also the social sciences and a wide variety of nonscientific (e.g. political, moral, literary, artistic, and religious) disciplines and practices. It is only by engaging in such inquiries, he thinks, that we can "get an adequate global view of the world, of thought, of language, or of anything" (Putnam 1962a: 41).

Acknowledgements

I presented a shorter version of this paper on September 18, 2016, at the Harvard University Philosophy Department's memorial conference titled "A Celebration of the Life and Work of Hilary Putnam". Thanks to Mario De Caro, Juliet Floyd, Warren Goldfarb, and Thomas Ricketts for their helpful comments on earlier drafts.

Cited works by Hilary Putnam

1962a, "The Analytic and the Synthetic," reprinted in Putnam 1975c, 33-69.

1962b, "It Ain't Necessarily So," reprinted in Putnam 1975b, 237-49.

1967, "Time and Physical Geometry," reprinted in Putnam 1975b, 198-205.

1975a, "The Meaning of 'Meaning'," reprinted in Putnam 1975c, 215-71.

1975b, *Mathematics, Matter, and Method: Philosophical Papers*, *Vol. 1*, Cambridge: Cambridge University Press.

1975c, *Mind, Language, and Reality*: *Philosophical Papers, Vol. 2*, Cambridge: Cambridge University Press.

2012, "Corresponding with Reality," in De Caro, M., and Macarthur, D. (eds.), *Philosophy in an Age of Science*, Cambridge, MA: Harvard University Press, 72-90.

2015, "Intellectual Autobiography," in Auxier, R.E., Anderson, D.R. and Hahn, L.E. (eds.), *The Philosophy of Hilary Putnam*, Chicago, Illinois: Open Court, 3-110.

Advisory Board