Contents

ARTICLES

REPORTS

Organon F 25 (4) 2018: 442-455

Some Remarks on the Mill-Frege Theory of Names

NICOLÁS LO GUERCIO[1]

ABSTRACT: In a recent paper García-Carpintero (2017) argues that proper names possess, in addition to their standard referential truth conditional content, metalinguistic descriptive senses which take part in semantic presuppositions. The aim of this article is twofold. In the first part I present an argument against García-Carpintero's presuppositional view, which I call the collapse argument. In short, I argue that the view has the unwelcome consequence of making contexts of use and *felicitous* contexts of use collapse. If this is correct, a presuppositional account of the metalinguistic descriptions allegedly associated with proper names proves incorrect. In the second part I sketch an alternative Millian strategy which is able to account for the evidence which allegedly supports the presuppositional view.

KEYWORDS: Pragmatics – presupposition – proper names – semantics.

1. The Mill-Frege theory of proper names

In order to understand García-Carpintero's theory (The Mill-Frege Theory of proper names) it is convenient to start by pointing at a number of assumptions on which such theory rests. The first one is the Kaplanian distinction between generic and specific names. A generic name consist just

[1] Received: 29 June 2018 / Accepted: 13 August 2018

✉ Nicolás Lo Guercio

Instituto de Investigaciones Filosóficas (IIF)
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
Pte. Luis Sánez Peña 1177, (1110), Ciudad Autónoma de Buenos Aires, Argentina

e-mail: nicolasloguercio@gmail.com

in a phonological or orthographical articulation, i.e. a pattern of sounds or ink-marks. Specific names, in turn, are individuated by an historical event, to wit, the act of naming by means of which the name was created (I'll say a little bit more about acts of naming below). Crucially, for García-Carpintero names that occur bare in argument position are *specific* names, not generic ones. The second assumption is Properism, roughly the view that all the objects named *John* possess different specific names, say $John_1$, $John_2$, $John_3$…, which share the same phonological and orthographical articulation.[2] Finally, García-Carpintero adopts a way of understanding the semantics/pragmatics divide according to which the semantic dimension is not restricted to the truth conditional realm but it comprehends "any meaning feature belonging to a type constitutive of the nature of languages (so that any attempt at characterizing a possible language having any chance of being the actual language of a population which overlooks that type of feature is thereby inadequate)…" (García-Carpintero 2000, 112).

With this in mind, let's now summarize the main points of the Mill-Frege view. First, the view is token-reflexive: concrete referents are ascribed to name-tokens, namely concrete actual or possible uses of expressions, as opposed to the more standard Kaplanian occurrences (cf. García-Carpintero 1998; 2000 for a discussion of the token-reflexive view). Second, the view is Millian in the following sense: it grants that the only truth conditional contribution of a name-token to the utterance of which it takes part is its referent. Thus, unlike some forms of descriptivism, the Mill-Frege theory does not claim that names are *synonymous* with definite descriptions, metalinguistic or otherwise. Still, proper names are *semantically* associated with descriptive metalinguistic senses, which figure in semantic presuppositions triggered by tokens of them. What takes us to the third point: the Mill-Frege theory maintains that a token **n** of a specific proper name N_i carries a semantic presupposition of the form 'x is the unique individual picked out in the act of naming instituting the N_i-appellative practice to which **n** belongs' (García-Carpintero 2017, 26). This presupposition is an instance of a schematic rule, which is said to be part of the linguistic knowledge of any competent speaker, of the form

---

[2] The view contrasts with Commonerism, viz. the position that all Johns share the same name.

> N_i = For any use n of proper name N_i, n refers to x if and only if x is the unique individual picked out in the act of naming instituting the N_i-appellative practice to which n belongs. (García-Carpintero 2017, 26)

García-Carpintero adopts a Stalnakerian view on presuppositions. Very roughly, on this view an utterance presupposes a proposition p if it is inappropriate unless p belongs to the common ground, that is, to the set of propositions commonly accepted by the conversational partners (otherwise it has to be accommodated by the audience). Stalnaker's view is considered a pragmatic view, usually opposed to the semantic, Strawsonian view on presuppositions. According to García-Carpintero, however, Stalnaker's approach is compatible with the existence of some presuppositions being semantic in the traditional sense (cf. García-Carpintero 2000 and 2016). Very roughly, an utterance (semantically) presupposes a proposition p if (i) the truth of p is required for the utterance to have a truth value and (ii) the presupposition is triggered by the conventional meaning of some expression in the sentence. On the Mill-Frege view presuppositions associated with proper names are semantic in exactly that sense: they are triggered by the conventional meaning of the name and their truth is required for the name to refer, hence for the utterance to possess a truth value.

We already said in which sense the theory is Millian. The last paragraph makes it clear in which sense it is Fregean: on the one hand, the view contends that names are *semantically* associated with descriptive senses, which are part of the linguistic knowledge shared by competent speakers; on the other hand, the view has it that descriptive senses *fix the reference*, i.e. they figure in semantic presuppositions the truth of which is required for the name to refer, thus for the utterance to have a truth value.

Let me end this brief summary by saying a few words on *acts of naming*. On the Mill-Frege theory acts of naming are 'purposeful events, instituting linguistic conventions, appellative practices' (García-Carpintero 2017, 16), thereby creating a new *specific* name and fixing a referent for it. Acts of naming can be explicit or implicit. In the former case, they constitute a directive speech act, a plea or request to create a new expression, i.e. to conventionally associate a given object with a generic name (a phonological or orthographical articulation) and to conform future practice to that convention. In the latter case the name comes into existence just by being

presumed to exist, provided that the community goes along with that presumption. Finally, acts of naming can be successful or not. When they are, a new *specific* name comes into existence, and the relevant object becomes the semantic referent of that name.

These are the main theses of the Mill-Frege theory. With these in mind I will discuss, in the next section, what I think is the main problem for this presuppositional approach to proper names.

2. The collapse argument

In line with the tradition (Karttunen 1974, Stalnaker 1974) we can think of presuppositions as restrictions on appropriate or felicitous contexts of use. Put differently, presuppositions restrict the contexts in which a sentence can be felicitously used to those in which the presupposition is satisfied: in the case of pragmatic presuppositions felicitous contexts are those in which the presupposition is part of the common ground; in the case of semantic presuppositions, felicitous contexts are those in which the presupposition is true. Thus, whenever an expression triggers a presupposition there is a set of contexts, viz. the ones in which the presupposition fails, in which using the expression would be inappropriate. By way of illustration, consider a presupposition typically thought of as semantic in nature, like the existence presupposition in

(1) The king of France is bald.

If I use (1) and there is no king of France, I have used the expression infelicitously. As a consequence, the story goes, the description does not denote and the utterance lacks a truth value. Something analogous can be said for the case of indexicals, at least on a presuppositional account (García-Carpintero 2000). On this view, these expressions possess a relational property involving any use, which is mutually known by hearer and speaker on the basis of linguistic knowledge alone, it is reasonably individuative and fixes the reference of the expression in a context. As in the case of proper names, this property is said to be part of a semantic presupposition associated with the expression. In the case of demonstratives, the presuppositional rule goes as follows:

> That = For any use that of *that*, that refers to x if and only if x is the unique entity (in a contextually specified class F) 'demonstrated' when that is produced. (García-Carpintero 2017, 10)

Here, again, if I use *that* without there being a unique entity demonstrated in the context I have used the expression inappropriately, and as a consequence my token does not refer and the utterance lacks a truth value.

If names trigger semantic presuppositions we expect the same behaviour: there should be some contexts in which I use a name but, since the presupposition is not satisfied, my use is infelicitous thus the name fails to refer and the utterance lacks a truth value.[3] However, there is a key difference between the case of definite descriptions and demonstratives and that of proper names. According to García-Carpintero 'each instance of N_i is a rule associated with a *specific* proper name: a word individuated by its linguistic features, *in particular the semantic one constituted by the act of naming which fixes its reference*' (García-Carpintero 2017, 26, my emphasis). Moreover, on the Mill-Frege theory of names, whether a given token **n** is a token of a certain specific name, N_i, is also determined by its linguistic features, in particular by the fact that it exploits the N_i-appellative practice instituted by a certain concrete historical act of naming which fixed its reference. Put differently, the property of being related with this or that original act of naming or appellative practice is individuative of the specific name N_i and tokens of it.

---

[3] At this point a clarification is needed. We expect the same behaviour from definite descriptions, demonstratives and proper names only as long as we assume (following García-Carpintero) that the three kinds of expressions carry a semantic presupposition. On that assumption, we expect presupposition failure to lead to an infelicitous use in the three cases. Now, although it is standard to associate a semantic presupposition to definite descriptions, it is less standard to do so in the case of demonstratives and proper names. Here, for the sake of the argument, I follow García-Carpintero in granting the existence of a semantic presupposition in the three cases in order to show, in the end, that proper names behave different both from (i) expressions commonly thought to carry a semantic presupposition, like definite descriptions and (ii) expressions not so commonly thought to carry a semantic presupposition, but which García-Carpintero himself would classify as involving semantic presuppositions, like demonstratives. Thanks to a reviewer for *Organon F* for signaling this point.

This point is clear on García-Carpintero discussion of Madagascar-like cases (García-Carpintero 2017, 18-20). There, he explicitly sides with Sainsbury (2005), and against Sainsbury (2015), in claiming that in Madagascar-like cases there is no change of reference of the same specific name, but the creation of a new appellative practice, hence of a new specific name. In other words, he rejects Sainsbury's (2015) claim that the historical chain that determines the same-name relation and that which determines the same-referent relation are different. In turn, he contends that

> this ignores a semantic constitutive role that the previous considerations show acts of naming to have. In a nutshell: they are intended to introduce a word; words are individuated in part by their semantic features; names, like indexicals, are *de jure* constitutively referential expressions, whose semantic referent is determined relative to what transpired at a particular act of naming. (García-Carpintero 2017, 20)

This raises a problem. If acts of naming and appellative practices are individuative of specific proper names and in addition, they are part of their presuppositional content, then presupposition failure, i.e. a token **n** not being associated with any act of naming/appellative practice, does not merely prevents the speaker from using the name felicitously but *it prevents her from using a specific name at all*, since the fact mentioned in the alleged presupposition is the fact which crucially determines whether such token is indeed a token of a specific name. To be sure, this is not the case for demonstratives in García-Carpintero's account: presuppositions associated with demonstratives play a role in fixing the reference, but they do not play any role in the individuation of the expression. The fact that there is a unique individual being demonstrated in the context is not part of the facts which determine whether the expression being used is this or that demonstrative. That's why, if such presupposition fails you have nonetheless used the demonstrative in question, although inappropriately. Something analogous can be said about the case of definite descriptions. There being a king of France plays no role in the individuation of the expression *the king of France*. Hence, you are able to use that very expression (although infelicitously) even if the presupposition fails. According to the Mill-Frege theory, in turn, being associated with a particular act of naming or being part of this or that appellative practice constitute the facts the obtaining of which

make it the case that a given token is in fact a token of a specific name. As a consequence, if the presupposition fails, i.e. if those facts do not obtain, the token you produced is not a token of a specific name at all; at most, what you did was tokening a generic name instead, viz. merely a token of a phonological or orthographical articulation.

In other words, if the descriptive metalinguistic sense in question figured in a semantic *presupposition* associated with the name, it should be possible to use the name infelicitously, that is, it should be possible to use that very name even though the presupposition fails. However, if we take the Mill-Frege theory seriously, it seems that it is not possible to do that: the Mill-Frege theory makes contexts of use and *felicitous* contexts of use collapse; all contexts of use of a specific name are contexts in which the alleged presupposition holds (otherwise, you wouldn't be using a name at all). If what has been said is correct, then the metalinguistic senses discussed by García-Carpintero are not part of a semantic presupposition.

Now, what about the evidence García-Carpintero provides in favour of the presuppositional view? The main argument consists in trying to show that proper names pass the 'Hey, wait a minute!' test:[4]

> We have been debating what to call the cat we recently adopted; I support 'Whiskers'. Other members of my household favour 'Flaubert'. A visit friend asks 'What is the new cat like?' Out of the blue, I answer:
>
> (2) Whiskers is adorable.
>
> A supporter of the alternative name promptly objects: 'Hey, wait a minute, I did not know we had agreed on calling it Whiskers!' (García-carpintero 2017, 15)

We can analyse the case in the light of the previous objection. García-Carpintero presents the Whiskers case as an example of an implicit act of naming. The speaker attempts at introducing a specific name for the cat, i.e. to make 'Whiskers' semantically refer to the cat. If the community accommodates the alleged presupposition the act of naming is successful and a new specific name comes into existence. If, in turn, the community

[4] I will discuss two other pieces of evidence presented by García-Carpintero in the next section, once I sketch the Millian strategy I favour.

refuses to conform to that practice, the act is nullified. But the question is, if the alleged presupposition fails, i.e. if the act of naming is nullified, has the speaker used (infelicitously) a specific name, *Whiskers*$_1$, which refers to the relevant cat?[5] If what I have argued in this section is correct, she has not. The speaker has tokened a generic name, the phonological articulation /Whiskers/, but she has failed to make a specific name out of it. Crucially, this is unlike typical cases of semantic presupposition: if the audience fails to accommodate the presupposition that there is a king of France after an utterance of (1) this does not nullify the fact that *the kind of France* is a legitimate expression of English.[6]

---

[5] García-Carpintero considers this question for explicit acts of naming of the form 'let's call … N'. He maintains that in those cases the occurrence of *N* is predicative and expresses a metalinguistic predicate whose application conditions involve a generic name. I agree with García-Carpintero on this (there is independent evidence that this type of occurrences are part of a small-clause involving a predicate – see Matushansky 2008). However, the question remains concerning implicit acts of naming (like the whiskers case) or simple occurrences of proper names in argument position in which the alleged presupposition is not satisfied. In this cases no expression of the form 'let's call … N' is used and the act of naming, if there is one, can only be implicit.

[6] A reviewer for *Organon F* brings up some possible counterexamples to my collapse argument. The first one is the case of empty names, like the famously discussed example of *Vulcan*. García-Carpintero does not discuss empty names in his paper, but we can apply the theory to this case. According to García-Carpintero's view, the sense associated with a token of a proper name, e.g. *Vulcan*$_i$, is something like 'whatever individual is picked out in the act of naming instituting the *Vulcan*$_i$-appellative practice to which **Vulcan** belongs'. This sense, in turn, is an ingredient of a semantic presupposition of the following form:

> For any use **Vulcan** of the proper name *Vulcan*$_i$, **Vulcan** refers to *x* if and only if *x* is the unique individual picked out in the act of naming instituting the *Vulcan*$_i$-appellative practice to which **Vulcan** belongs.

The reviewer interprets the right hand of the bi-conditional as involving an existence presupposition, something like '$\exists x$ such that *x* is the unique…' so that the presupposition fails if and only if no unique object was picked out in the act of naming. So, since there is no unique individual picked out in the act of naming instituting the *Vulcan*$_i$-appellative practice, the presupposition fails, hence a token of *Vulcan*$_i$ would be an infelicitous use of the name *Vulcan*$_i$. If this is correct, there are cases in which we see no collapse.

I have been interpreting the view differently. On my interpretation, the satisfaction of the presupposition does not guarantee that the name has a referent. This is suggested

To sum up. Presuppositions are constrains on felicitous contexts of use. So if a name triggers a presupposition you expect there'll be some felicitous uses of the name and some infelicitous uses thereof, depending on whether the presupposition is satisfied. But this is not what we see in the case of proper names (e.g. in the Whiskers case): if the presupposition is satisfied you have a felicitous use of the name; but presupposition failure nullifies the act of naming, preventing the name coming into existence. In that case you do not have an infelicitous use of the name, because you do not have a use of a name at all. Since this is not the behaviour one expects from semantic presuppositions, we must conclude that the metalinguistic description in discussion is not a semantic presupposition.

---

by García-Carpintero's treatment of the Whiskers case, one of the examples of presupposition failure he discusses. In that case, the problem is not that the name does not refer, but that the act of naming failed. It is also the more charitable interpretation of the view if one considers *Vulcan* examples in the light of the 'Hey, wait a minute!' test:

A: Vulcan is a planet

B: #Hey, wait a minute! Vulcan does not refer to anything!

If my interpretation is correct, however, the alleged presupposition associated with the name is in fact satisfied in *Vulcan* examples: the token of the name in fact refers to whatever individual is picked out in the original act of naming (since there is no individual picked out in the act of naming instituting the $Vulcan_i$-appellative practice to which **Vulcan** belongs, **Vulcan** does not refer).

Secondly, the reviewer suggests that if one's metasemantics of 'semantic reference' for uses of proper names includes not only the intentions of the speaker but also the availability of such intentions to the audience, there might be infelicitous uses of proper names too. As I understand the suggestion, those would be uses in which a token of a name is appropriately related with a certain appellative practice and the speaker uses the name with the intention to refer to the unique individual fixed by that practice, but such intentions are not available to the audience. However, it is important to point out that this would constitute a non-trivial departure from García-Carpintero's view. Specifically, it would introduce requisites concerning the intentions of both speaker and audience into the presuppositions associated with proper names. Maybe this alternative view escapes my collapse argument, but (i) the view still has to be shown to be plausible and (ii) I don't think these cases threat the collapse argument as an argument against García-Carpintero's current version of the Mill-Frege theory.

3. A Millian story

According to the Mill-Frege theory, an act of naming creates a *specific* name, N_i, and fixes a referent for it, thereby instituting an appellative practice related to that name. A certain token is a token of N_i only insomuch as it exploits that appellative practice, in which case it refers to whatever object is fixed by the original act of naming, contributing only that object to truth conditions. Now, García-Carpintero maintains that the aforementioned facts figure in the semantics of the name, specifically as part of its presupposed content. In this section I sketch a Millian proposal which accounts for the aforementioned facts without making them part of the semantics of the name (not even in the broad sense of 'semantics' adopted by García-Carpintero).

In a Millian framework one can think of acts of naming and appellative practices as a collection of non-semantic (social, psychological, causal) facts[7] which ground a number of semantic facts: (i) the fact that there exists a certain specific name N_i, which possess a certain semantic reference—in the Kaplanian vocabulary, a constant character which outputs the same object, i, in every context—and is conventionally associated with a certain phonological string /N/, and (ii) the fact that a particular token of the phonological string /N/ is in fact a token of N_i, which inherits its linguistic features. Within this view, the mere fact that a token **n** is produced *as a token of the specific name* N_i, conveys (in some sense of 'convey' to be specified) the information that the previously mentioned social, psychological and causal facts obtain. In other words, it conveys the information that an appellative practice is in force, traceable to an original act of naming.

In order to make sense of this view, it is crucial to find a suitable way of understanding 'convey' in the previous paragraph, in particular one that

[7] A thorough treatment of the numerous facts which determine the semantic reference of a proper name is beyond the reach of this article. It will suffice to note that it will involve psychological facts, like the existence of certain intentions on the part of language users, both in acts of naming which create proper names as in referential uses which conform to the practice, sociological facts, in order to account for what Sainsbury (2005, 106) calls *unwitting* baptisms (i.e. events which lead to an unintentional creation of an appellative practice) and causal facts (consider again Sainsbury's view, according to which some baptisms require an object-related intention/mental state, for which a causal interaction with the object is required).

does not involve incorporating the aforementioned facts as part of the semantics of the name. We can find a good candidate in Predelli's idea of use-imparted information (Predelli 2013; 2017), viz. information imparted by virtue of extra-semantic regularities encoded in the use of an expression. Predelli cashes out this notion in terms of what he denominates *settlement*. There are different forms of settlement, but the relevant in this context is Mere Settlement:

> Mere Settlement
>
> A use of an expression e in a context c *merely settles*^{T} a sentence S iff whenever $c \in CU^T(e)$, S is True.[8] (cf. Predelli 2013, 32-34)

That is, a use of an expression e merely settles^T a sentence if and only if that sentence is true in every context of use of e of a certain type T.[9] By way of illustration consider some of Predelli's examples. Since arguably tokening a linguistic expression requires intentional agents, every use of an expression in a context c merely settles the sentence 'there exist, have existed or will exist intentional agents in c_w'. That is, in every context of use of any expression the foregoing sentence is true. Likewise, every use of 'I am hungry' settles 'there exists, have existed or will exist tokens of a sentence which contains an indexical'. Crucially, this information is not conveyed by semantic means, wherever you draw the semantic line: it is information imparted by virtue of the peculiarities of linguistic use.

Now, the Millian can think of the kind of metalinguistic information which García-Carpintero locates at the presuppositional level as use-imparted information, merely settled by uses of proper names. According to this idea, a context of use of a specific name N_i is a context in which there is a token **n** of N_i, endowed with a particular character, viz. a constant function which outputs an object i in every context. Now, as we already said, on this view tokens are individuated by their belonging to a certain appellative practice and their being related with an original act of naming which uniquely fixed its semantic referent. Hence, every context of use of N_i is a

---

[8] Every sentence which is true in virtue of character alone is settled by the use of any expression. Mere settlement leaves aside these sentences.

[9] Some sentences can be merely settled for some types of use but not for others. However, some sentences can be merely settled for *all* types of use.

context in which at c_w at least one expression semantically refers to i, that is, a context in which there is at c_w a corresponding act of naming/appellative practice which conventionally relates the generic name /N/ to i, so that i is the unique individual picked out in the act of naming instituting the N_i-appellative practice to which **n** belongs. In other words: every use of a specific name imparts the information that the non-semantic facts which must obtain for the specific name to exist, in fact obtain. Crucially though, this is not information encoded in the conventional meaning of the name but conveyed in virtue of the fact that a *specific* name was used.

The view roughly sketched above has the virtue of explaining several facts discussed by García-Carpintero, which supposedly support the thesis that there are metalinguistic senses which belong to the semantics of proper names. One of these facts concerns some patterns of inference which appear to be problematic for Millianism:

Peter is hungry
∴ Someone called Peter is hungry

Although logically invalid, this inference seems acceptable in some circumstances (see Leckie 2012 and Schoubye 2016). This represents a challenge for the Millian: it is not possible to account for the acceptability of the inference in terms of truth conditional content; but it seems that truth conditional content is the only explanatory tool the Millian has at her disposal, since for her names are just tags whose sole semantic contribution is the object to which the name refers. The Mill-Frege theory, in turn, can offer a straightforward explanation: the inference is not valid in general, but it is truth-preserving across felicitous contexts, i.e. contexts in which the presupposition of the premise is met and the premise is true. However, the Millian view previously sketched can also account for these patterns. On this view every use of a proper name N_i *merely settles* the sentence 'someone is called N_i'. Hence, the inference is truth-preserving for every context in which the premise is *used* and is true.

Another piece of evidence that García-Carpintero presents as supporting the Mill-Frege view has to do with speakers' awareness of the metalinguistic description in question, manifested in their disposition to accommodate alleged presuppositions (as it would be the case in the Whiskers case, if the audience did not object the assertion). This fact, however, is

compatible with a Millian account as the one presented above. On this view the existence of a social convention (an appellative practice) which is being exploited by the name user is information imparted by extra-semantic means, by virtue of it being the fact which individuates the name and grounds its having these or those semantic features. Thus, it is expected for speakers to be able to somehow recover or become aware of the fact that an appellative practice is in place or, to the contrary, to point out that the token produced is not associated with any appellative practice, i.e. it is not a token of an actual specific name.

In addition, we should point out that accommodation alone is not a reliable test for presuppositionality. Accommodation involves adjusting the 'conversational score', in Lewis' terms, so as to ensure (ceteris paribus and within certain limits) that the speakers' utterances will come out true, or otherwise acceptable (Lewis 1979). The process might take place in the face of presupposed content but also while fixing the values of implicit arguments, establishing reference points ('come', 'go' and so on) or, crucially, resolving lexical or structural ambiguity. So the fact that the audience is ready to accommodate the proposition that the referent of a certain name is called 'N' is compatible with the non-presuppositional, Millian position just outlined: if someone uses an articulation /N/ bare in argument position the audience will try to ensure (ceteris paribus and within certain limits) that the utterance will come out true or otherwise acceptable. That involves accommodating the fact that the articulation employed is that of a specific name, i.e. that there is an appellative practice which is being exploited by the speaker. Put differently, the fact that the audience accommodates is not the semantic fact that the specific name being used refers to this or that object, but the pre-semantic fact that the articulation being tokened is in fact a specific name.

4. Conclusion

To sum up, we have shown that the Mill-Frege view faces a serious problem. On the one hand the claim that names trigger semantic presuppositions involving a descriptive metalinguistic sense, when combined with the theory's stance concerning the way of individuating specific names, has an unwelcome consequence, namely that of making contexts of use and

felicitous contexts of use collapse, against presuppositions' expected behaviour. On the other hand, I showed that several facts which allegedly support the presuppositional view can be accounted for within a Millian approach.

Acknowledgments

I want to thank all the people of the Buenos Aires Linguistics and Philosophy of Language group, especially Eleonora, Orlando, Andrés Saab and Ramiro Caso for discussion on previous versions of this work.

References

GARCÍA-CARPINTERO, M. (1998): Indexicals as Token-Reflexives. *Mind* 107(427), 529-564.

GARCÍA-CARPINTERO, M. (2000): A Presuppositional Account of Reference Fixing. *The Journal of Philosophy* 97(3), 109-147.

GARCÍA-CARPINTERO, M. (2016): Accommodating Presuppositions. *Topoi* 35(1), 1-8.

GARCÍA-CARPINTERO, M. (2017): The Mill-Frege Theory of Proper Names. *Mind* Online: https://doi.org/10.1093/mind/fzx010

KARTTUNEN, L. (1974): Presupposition and Linguistic Context. *Theoretical linguistics* 1(1-3), 181-194.

LECKIE, G. (2013): The Double Life of Names. *Philosophical Studies* 165(3), 1139-1160.

LEWIS, D. (1979): Scorekeeping in a Language Game. *The Journal of Philosophy*, 339-359.

MATUSHANSKY, O. (2008): On the Linguistic Complexity of Proper Names. *Linguistics and Philosophy* 31(5), 573-627.

PREDELLI, S. (2013): *Meaning without Truth*. Oxford: Oxford University Press.

PREDELLI, S. (2017): *Proper Names: A Millian Account*. Oxford: Oxford University Press.

SAINSBURY, R. M. (2015): The Same Name. *Erkenntnis* 80(2), 195-214.

SAINSBURY, R. M. (2005): *Reference without Referents*. Oxford: Clarendon Press.

STALNAKER, R. (1974): Pragmatic Presuppositions. In: Munitz, M. K., & Unger, D. K. (eds.), *Semantics and Philosophy*, vol. 80. New York University Press.

SCHOUBYE, A. J. (2016): Type-Ambiguous Names. *Mind* 126(503), 715-767.

Organon F 25 (4) 2018: 456-472

Conciliationism and Fictionalism

MARCUS WILLIAM HUNT[1]

ABSTRACT: This paper offers fictionalism as a new approach to the problem of reasonable disagreement discussed in social epistemology. The conciliationist approach to reasonable disagreement is defined, and three problems with it are posed: that it is destructive of inquiry, self-defeating, and unacceptably revisionary. Hans Vaihinger's account of fictions is explained, and it is shown that if the intellectual commitments that are the subject of reasonable disagreements are treated as fictions rather than as beliefs, the three noted problems are avoided. Whereas beliefs have a "rivalrous" relation to the source of their justification (evidence), fictions have a non-rivalrous relation to the source of their justification (expediency), meaning that disagreement over which fictions to employ is not problematic in the way that disagreement over what to believe is. Some objections to the fictionalist approach to reasonable disagreement are answered.

KEYWORDS: Conciliationism – disagreement – fictionalism – pessimistic induction – social epistemology – Hans Vaihinger.

0. Introduction

In this essay I address three criticisms regarding the consequences of the conciliationist approach to disagreement; that conciliationism is destructive

[1] Received: 15 June 2018 / Accepted: 24 August 2018

✉ Marcus William Hunt

Department of Philosophy, School of Liberal Arts, Tulane University
105 Newcomb Hall, 1229 Broadway, New Orleans LA 70118

e-mail: mhunt4@tulane.edu

of inquiry, self-defeating, and unacceptably revisionary. Although I endorse the doxastic revision that conciliationism suggests, my intention in this paper is not to defend conciliationism, but rather to draw a new approach toward reasonable disagreement from the examination of these three criticisms. I suggest that the doxastic revision which conciliationism requires of us is consistent with our maintaining a propositional attitude other than belief towards our intellectual commitments. I suggest that we ought to relate to them as fictions, and that doing so avoids the three noted criticisms. In part one I briefly state what I take conciliationism to be. In part two I explain the three criticisms of conciliationism. In part three I explain Hans Vaihinger's account of fictions. In part four I show how fictionalization of our intellectual commitments avoids the three criticisms. In part five I state and respond to five objections to my suggestion.

1. Specifying conciliationism

"Conciliationism" and "the steadfast view" are the two approaches to the question of the extent to which doxastic revision is rational for the disputants of a reasonable disagreement. The former calls for substantial doxastic revision, and the latter calls for little or no doxastic revision. One way of stating the insight of conciliationism is that in cases of reasonable disagreement one is provided with second order evidence which weighs against the first order evidence one holds for one's belief or undercuts this evidence.

On conciliationism, not all disagreements give us reason for revising our beliefs, only reasonable disagreements do. For there to be a reasonable disagreement, the disputants must be epistemic peers. Epistemic peers are those who meet these two conditions:

> "(i) they are equals with respect to their familiarity with the evidence and arguments which bear on that question, and
> (ii) they are equals with respect to general epistemic virtues such as intelligence, thoughtfulness, and freedom from bias" (Kelly 2005, 175).

I will make two controversial assertions about these conditions which allow conciliationism to apply to many of our real-world disagreements,

rather than remaining a doctrine that applies only to model cases of disagreement in which we merely stipulate that (i) and (ii) obtain.

First, "equal" need not mean "same." To say that two disputants are epistemic peers with respect to (i) and (ii) is not to say that they have the very same evidence or epistemic virtues. If sameness of (i) and (ii) is required for a reasonable disagreement, then it seems unlikely that anyone has ever had a reasonable disagreement, given the uniqueness of every individual's life experiences (King 2012). But real-world disagreements are, surely, sometimes reasonable disagreements. So, a conception of "equal" other than sameness should be adopted. I suggest a dialectical understanding of "equal." That is, so long as the disagreement of the two disputants could not (counterfactually) be resolved by a mutual disclosure of their various dissimilarities, there is a reasonable disagreement. The extra tidbit of evidence that disputant A has only makes A and B not peers if this evidence ought to significantly change the beliefs of disputant B if it were disclosed to B. That A has reviewed 1001 case studies and B has reviewed 1000 case studies does not make A and B non-peers, nor does that A is a little more open minded whilst B is a little more attentive.

Second, note that the definition of "epistemic peers" offered is metaphysical. Yet, as an epistemological matter, in a given real-world disagreement we are often not quite sure whether the disputants are epistemic peers or not. In real-world disagreements we are often locked in "apparently" reasonable disagreements, where it seems that both (i) and (ii) may well obtain. It seems that we have grounds for some degree of doxastic revision not only when we know that (i) and (ii) obtain, but so far as it seems they may obtain. Where the disagreement is clearly attributable to a lack of evidence or failure of reason in one party, the disagreement gives one no grounds for revising one's beliefs, but where the disagreement cannot be attributed to such factors, the hypothesis that the disagreement is a reasonable one remains plausible and provides grounds for some degree of doxastic revision.

2. Three criticisms of conciliationism

I now turn to stating the three objections to conciliationism that I wish to address.

2.1. Pessimistic induction

The first criticism is that the doxastic revision advised by conciliationism, even though epistemically required, would put an end to much intellectual inquiry. As a conciliationist, one finds oneself in a somewhat exotic form of the pessimistic induction problem (Laudan 1981) that counsels against belief in the theoretical entities posited by contemporary science, that is, against scientific realism. Surveying the history of philosophy or social science one would find, on the characterization of epistemic peerhood I have offered, that most of the disputant's disagreements were apparently reasonable disagreements. Therefore, the peers to these disagreements ought to have conciliated and become adoxic (rather than maintaining, for example, nominalism and realism, Keynesianism and monetarism). Likewise, one finds oneself in apparently reasonable disagreements, in many cases over the same or similar matters as past thinkers. Therefore, by induction one has good reason to believe that in 30 years, when the next iteration of philosophy and social science has emerged, one will again find oneself in apparently reasonable disagreements about the same or similar questions – again requiring conciliation. Therefore, one reasonably anticipates never having justified beliefs about these matters. If one's end in philosophizing is to hold justified beliefs about philosophical matters (or to hold any other type of mind-to-world direction of fit propositional attitude that is sensitive to evidence), then by ought implies can one should not bother philosophizing. One should become dispirited with philosophizing, and go to tend the garden, or some activity about which one can reasonably anticipate a decent chance of success. Note that this pessimistic induction is more severe than the one facing scientific realism, since there is at least some chance that present scientific theories will not succumb to the same fate as those of the past (being superseded) whereas in this case we know that the problem (apparently reasonable disagreement) already obtains for our present beliefs. Likewise, the various strategies for defending scientific realism, such as selective realism (Hardin & Rosenberg 1982; Psillos 1999), entity realism (Cartwright 1983; Hacking 1982), structural realism (Worrall 1989), or claiming that the scientific theories of the present are "more successful" than those of the past or in some way qualitatively different (Fahrbach 2011), seem hard to replicate *vis-à-vis* philosophical and social scientific theories or to not really admit of analogues at all.

2.2. Self-defeat

Conciliationism has been charged with being self-defeating (Christensen 2009, 762; Plantinga 1995). Conciliationists recommend substantial doxastic revision in the light of apparently reasonable disagreement. Steadfasters instead recommend little or no doxastic revision in the light of an apparently reasonable disagreement. Plausibly, steadfasters are in an apparently reasonable disagreement with conciliationists. According to conciliationism this demands that one undergo a substantial doxastic revision away from conciliationism. This effect can be iterated *ad infinitum* in very messy and contradictory directions, e.g. "Cynthia conciliationist" conciliates with the steadfaster to some intermediate position, but in turn meets "Cal conciliationist" who thinks he has a special reason not to conciliate with the steadfaster. It appears to Cynthia that she is in an apparently reasonable disagreement with Cal, so she conciliates closer to the original conciliationist position. In a world of recursive debates, maintaining the sort of "50:50" adoxicism typically suggested by conciliationism would rarely be justified.

2.3. Revisionary

Conciliationism seems too revisionary and counter-intuitive. We can imagine the case of a philosopher who has spent many decades carving out theories, and has many intuitions, about various matters: the immorality of abortion, the truth of pansychism, etc. Then they hear about conciliationism, find it to be convincing, and now have to abandon their many beliefs. From a reflective-equilibrium point of view, conciliationism is a case of the tail wagging the dog. If one must weigh one's theories and intuitions about all these other matters against one's theories and intuitions about the epistemology of disagreement, this latter must lose out.

3. Fictions

In *The Philosophy of "As If"* Vaihinger distinguishes between two types of ideation; hypothesis and fiction. Typically, we are under the impression that all of our intellectual commitments fall under the former category. Opinion, belief, and knowledge may be counted under the umbrella of

"hypothesis" because they share that they are intended to correspond to the world. Vaihinger's claim is that some of our intellectual commitments are not of this kind. Rather, there is a class of ideations called fictions which may be characterized as "products of the imaginative faculty" (Vaihinger 1935, 63). These are not intended as claims about reality, and are "advanced with the consciousness that [they are] an inadequate, subjective and pictorial manner of conception, whose coincidence with reality is, from the start, excluded" (Vaihinger 1935, 268). According to Vaihinger, fictions are representations which are known to be false or impossible. They induce us to think as if something that is the case were not the case, or as if something which is not the case were the case. The reason we should be interested in entertaining such ideations is that doing so proves to be useful in the wider process of theorizing or in practical activity. Despite their falsity, fictions "remain from a practical standpoint necessary elements in our thought" (Vaihinger 1935, 134). Therefore, whereas the justification for a hypothesis is evidence, the justification for a fiction is its expediency. Expediency here I will characterize loosely as that which aims at any good not immediately and narrowly concerned with corresponding to reality in the way that "hypothesis" is; guiding action, organizing thought, generating new hypotheses, regulating emotions.

According to Vaihinger, fictions play a role in many aspects of our intellectual and practical lives. I will mention a few examples for the purposes of illustration. In economics we create models which could never obtain in reality; a market in which there is perfect information, homogeneous products, no barriers to entry or exit, etc. Similarly, we may think about impossible utopias to help draw normative conclusions about what we should do. To the same end we might try to reason as if we were behind a veil of ignorance (Rawls 1999, 118-123). I may try to think about some matter as if I were you. We may import analogies from one field of thought to another; we may think about society as if it were an organism or a family, or about economic competition as if it were a process of Darwinian evolution. We may instruct a jury to deliberate as if they did not know a piece of excluded evidence that they do know. We may treat a human being in a permanent vegetative state, or a newborn, as if they were persons.

My suggestion is that we treat those of our intellectual commitments subject to reasonable disagreement as fictions. Whereas Vaihinger asserts that fictions are imaginative representations that are known to be false or

impossible, it seems that a weaker characterization captures what is essential; that fictions are imaginative representations not intended to correspond with reality (but where a possible correspondence with reality is not "from the start, excluded"). A good example here is free-will; an ideation that it seems can be consciously employed as a fiction in much juridical and ethical practice and reasoning whilst deliberately not taking a stand on the question of whether free-will as a metaphysical hypothesis is known to be false or impossible, or even true.

Conciliationism does not advise us to believe that our intellectual commitments are false. But it does give us grounds for not believing them and (with the addition of the pessimistic induction) for anticipating that we will not be able to justifiably believe them for the foreseeable future. The consciousness of this seems enough to sustain us in the practice of relating to our intellectual commitments as fictions rather than as hypotheses: images of how it is useful to think about the world regardless of what the world is really like. Note that relating to one's intellectual commitments as fictions does not completely foreclose the possibility that it may at some time become legitimate to regard these intellectual commitments as hypotheses, but it offers an alternative way of relating to them until such a time may come. For example, we might imagine someone who undergoes an intellectual journey in which they first relate to the idea of God as a hypothesis ("Since there is evil, God does not exist"), then as a fiction ("It is good to act as if God exists", "Investigate nature as if it were an orderly production"), then as a hypothesis again ("I've reconsidered the ontological argument – God exists!"). We might denote these different ways of relating to an intellectual commitment with a subscripted h or f, e.g. God_h or God_f.

4. Solutions to the three criticisms of conciliationism

I now explain how treating our intellectual commitments as fictions rather than beliefs resolves the three noted criticisms.

The key point pertaining to all three criticisms is to note a difference between beliefs and fictions in the manner of their justification. Because beliefs are intended as claims about reality then, out of a set of mutually contradictory beliefs, only one belief about some matter can be true. One is justified in holding a given belief rather than one of its competitors by

the evidence one has. And, typically, evidence counting for one belief is evidence against its competitors. For instance, evidence that supports the belief "Elvis is alive" counts against the belief "Elvis is dead." This is to say that beliefs are "rivalrous" – they compete for the justificatory resource of evidence (like rival gold-mining crews) and for one belief to become better justified by the discovery of new evidence typically means a depreciation in the justificatory value of the evidence for another belief (like the devaluation one such crew may inflict on another by finding the mother lode). In the case of apparently reasonable disagreement, that someone meeting (i) and (ii) holds different beliefs than oneself functions, according to the conciliationist, as a piece of evidence against one's own beliefs. By contrast, fictions do not make claims about reality and fictions are justified by their expediency not by any relation they bear to evidence. Whereas only one belief out of a set of mutually contradictory beliefs can be the best (the true), there seems to be no reason for assuming that there must be one fiction which is uniquely the best (the most expedient) in every respect and every context. Again, that one belief has had a lot of evidence adduced for it tends to show that it is the best (the true), but that some fiction has proven expedient does not tend to show that it is the best (the most expedient). In this way, fictions that are justified by expediency are "non-rivalrous." That someone employs a different fiction to the fiction that I employ, and does so to great effect, does not by itself show that my fiction is inexpedient. Concretely, if one political scientist investigates political institutions using a model that treats politicians as if their only motivation was to hold office and finds this fiction to be very expedient (in generating testable hypotheses, directing new research, etc.), this does not show that some other fiction would be inexpedient, e.g. a model that treats politicians as if their only motivation was to make money. By contrast, the more evidence ascertained for the belief that the only motivation of politicians is to hold office, the less reasonable it becomes to hold competing beliefs.

I anticipate that two difficulties will be raised. First, whilst the great expediency of one fiction may not act to "defeat" whatever expediency another fiction may have, and does not provide grounds for refusing to try-out some new fiction, it does provide *prima facie* reason to shift from employing some less expedient fiction to employing the more expedient fiction. Concretely, there may be some expediency to a Marxist approach to political economy, but there is perhaps greater expediency in a neo-classical approach

to political economy. Even if the expediency of the one does not detract from or destroy the expediency of the other, one ought to employ the latter. As Vaihinger himself remarked; "Expediency not only determines the acceptance or rejection of a particular fiction but also its selection from among others" (Vaihinger 1935, 90). Second, it seems that epistemic peers may reasonably disagree in their *beliefs* about which fictions it is most expedient to employ, leading to adoxicism about this question, meaning that they will have to suspend judgment about which fiction to employ, meaning that the dispiriting effect of conciliationism has not really been avoided.

In response to both these difficulties, note a distinction between the monolithic and the ecological expediency of a fiction. The former refers to what the most expedient fiction is when considered in isolation, i.e. a scenario in which the intellectual community had to collectively choose to adopt one approach to political economy or the other. The latter refers to what the most expedient combination of fictions for an intellectual community might be, i.e. whether it is more expedient that everyone agree on one approach to political economy, or whether it is better that several approaches are employed – and if so which approaches. The two noted difficulties are only difficulties for the monolithic expediency of a fiction. In the ecological sense of expediency, that someone else's use of a fiction has proved very fruitful for their research does not always provide one with a reason to shift to employing that fiction. For instance, a particularly expedient fiction may have attracted numerous researchers, such that one's net input to the intellectual enterprise is greater if one focuses on what is an overall less expedient but underutilized fiction; for instance, one might deliberately adopt a strange cousin of rational choice theory rather than rational choice theory itself. Or again, one may have a different purpose in mind for which some other fiction may be more expedient. Regarding the second difficulty, that many people are locked in apparently reasonable disagreements about which the most expedient fiction is in the monolithic sense ("Which theory is the best"), although it does provide grounds for doxastic revision about this question, it does not provide reason for hesitating in the selection of a fiction, since it is generally not true that expediency will be maximized by everyone employing the same fictions. Whilst there are numerous benefits to a shared paradigm (related to as a fiction or not), it is expedient that people choose to employ different fictions, even

ones which we may regard as unpromising. For instance, economics is plausibly a more progressive discipline because of the coexistence of Austrianism, Keynesianism, Behavioralism, etc., and the struggles between them than it would be if every researcher held the same intellectual commitments. Reasonable people can be aware of this, and so will not view differences in belief about which fictions it is most expedient to employ as a reason for indecision in employing a given fiction; any one of a number of fictions will remain reasonable options. Indeed, one may have nothing more than subjective grounds for employing a fiction ("it seems plausible to me," "it seems like a good approach"), or arbitrarily adopt a fiction on a volitional basis, but nevertheless find expediency in employing it.

This is not to say that the selection of any fiction in any circumstance will be expedient in the ecological sense. Even taking into account the synergy that results from allowing different forms of thought to flourish, compete, and cross-pollinate, certain fictions are so evidently inexpedient (or may become so with the passage of time) that one would reasonably regard someone who selected them as no longer being a peer, e.g., a political scientist who investigates political institutions using a fictional model that treats politicians as if their only motivation was to acquire letterheaded paper. On an analogy with Feyerabend's epistemological anarchism (Feyerabend 1975), we might call this view "fictional minarchism." The conditions under which it becomes expedient to begin employing some fiction or other can likely not be put in a general formula, and if they could it would be an extremely difficult to ascertain when such conditions obtained as an empirical matter. No doubt practical wisdom is called for. For instance, I would hazard that there is ecological expediency in the existence of small groups of creationists, phrenologists, mercantilists, and Steady State theorists, if only insofar as they inadvertently help clarify the commitments of researchers within the mainstream paradigms.

4.1. Pessimistic induction

Fictionalization of our intellectual commitments means that they avoid the pessimistic induction. Since fictions are not assertions about reality, it is no objection that previous intellectual commitments turned out to be false. Again, since fictions are justified by their expediency rather than any relation they bear to evidence, that intellectual commitments past, present,

and future, have a weighty piece of second order evidence going against them is not a relevant ground for the revision of one's fictions.

A different sort of pessimistic induction, grounded in the general inexpediency of many or most past fictions might give one reason to expect that one's fictions will prove inexpedient. Since expediency is a comparative concept only an observation of a trend of decline in the expediency of fictions employed would give one reason for doubting the expediency of one's own. But I take it that the real historical record of philosophy and the social sciences does not support this; whether our intellectual commitments in this area have proven expedient at all may be questionable, but the case for a gradual degeneration in their expediency seems hard to make.

4.2. Self-defeat

Conciliationism, as originally conceived, as well as being an epistemological or methodological claim about what we ought to believe under conditions of apparently reasonable disagreement, is itself offered as an object of belief. Such a conciliationism$_h$ is indeed self-defeating. But we can also endorse conciliationism$_f$, a fiction that concerns the most expedient way of relating to one's intellectual commitments under conditions of apparently reasonable disagreement. Conciliationism$_f$ could be characterized as "treating apparently reasonable disagreements as if they provide reason for conciliation." That others endorse steadfast$_h$ or steadfast$_f$ ("treating apparently reasonable disagreements as if they provide no reason for conciliation") is now not a reason for someone who endorses conciliationism$_f$ to conciliate about conciliationism$_f$; since the differences in approach between the two fictions are not a disagreement of theoretical reason but a difference in practical reason conciliationism$_f$ is not self-defeating.

The question then might seem to be whether conciliationism$_f$ or steadfast$_f$, or something else, is the most expedient fiction. Here, it might seem that steadfast$_f$ is the most expedient fiction; continuing along as if "I am right and others are wrong." Steadfast$_f$ allows one to practically ignore the dispiriting conclusion of the pessimistic induction. Moreover, it might at first seem that conciliationism$_f$ is a very inexpedient fiction, since it advises detaching from all of one's intellectual commitments, as if they were all, and would remain, subject to a weighty piece of undercutting evidence.

Although these characterizations of steadfast$_f$ and conciliationism$_f$ seem right, attempting to figure out which approach is the better and then adopting it is misconceived, given the previous remarks about ecological expediency. There is therefore likely to be room in the intellectual community for personalities who adopt either steadfast$_f$ or conciliationism$_f$, and there is call for an individual switching between them depending on the inquiry being pursued. Judgments about when it is best to adopt either of these fictions about one's intellectual commitments are no doubt difficult and highly contextual. At any rate, the self-defeat objection to conciliationism$_h$ is avoided by its fictionalized analogue.

4.3. Revisionary

Conciliationism$_f$ is not a hypothesis to be believed, but a fiction. It therefore makes no attempt to be consistent with the evidence one has for one's favored theories or with the evidence of intuition. Rather, to cite these things as grounds for not endorsing conciliationism$_f$ would be a category error, because fictions draw their justification from their expediency not from any relation they bear to evidence.

Having outlined my resolution of these criticisms, I now address five objections, in part as a clarificatory exercise.

5. Objections answered

5.1. "People believe their intellectual commitments and don't treat them as fictions. People cannot think like that."

I take it that the first sentence of this claim is for the most part descriptively accurate, but not worrying. The second sentence of this claim would be worrying if it is true. If it is true, fictionalization of our intellectual commitments would remain justified in principle. But there would be an "ought implies can" problem, and a certain frivolousness, in recommending that people think in a way that they are unable to think. In response, I would suggest that it is descriptively more precise to say that people cannot relate to their intellectual commitments as fictions all or most of the time, rather than that people are unable to do so at all. In this respect, there are many philosophical companions in guilt; skepticism

about causation, solipsism, and so forth, can only be sustained with effort for a short time before by "a kind of laziness…I happily slide into old opinions" (Descartes 1996, 15). The same difficulty seems to attend conciliationism itself, so in this respect my suggestion is no worse off than the view from which it departs.

It seems that the psychological possibility worry can also be addressed by pointing to numerous cases of philosophers and other inquirers explicitly treating very important elements of their thought as fictions; Thomas Hobbes' state of nature (Hobbes 1998, 85), John Rawls' veil of ignorance (Rawls 1999, 118-123), David Hume's account of justice and property (Hume 2007, 316-317), Edward Craig's account of knowledge (Craig 1991), Bernard Williams' account of truthfulness (Williams 2002, 20-22), Friedrich Forberg on God, freedom, and immortality (Forberg 2010), the models of economists, the rational choice theory employed in political science, the domestic analogy of international relations, teleology in biology, the legal treatment of rivers as persons, etc., each of which is likely to outrage or bemuse any undergraduate who mistakes them as hypotheses to be believed.

Further, although it seems most inquirers hold firm beliefs about their area of inquiry, much of the language they use suggests otherwise. One often hears an academic refer to their ideations as "projects," "research programs," or "orientations" (Hayek 1955, 225), which may "work out" or allow one to "tell that story" or "make that move." Much intellectual activity might comfortably and charitably be reinterpreted as fictive. For instance, a Marxist anthropologist might say "When I examine a society previously unknown to me, I do so as if each feature of its religion, morality, and law, was explained principally by the society's mode of production." One thing they might be doing is employing Marxism as a hypothesis to explain this society, and also seeing if Marxism as a hypothesis is falsified by the evidence this new society gives. But another thing they might be doing is self-consciously treating Marxism as an entirely unfalsifiable fiction that is expedient at gaining certain insights into this society or organizing inquiry about it. We are accustomed to condemning the Marxist anthropologist's claim for being unfalsifiable, but the real ground of its condemnation might be its inexpediency (if it is inexpedient).

5.2. "If we think of our ideations as fictions, we will stop caring about them."

Given the examples related above and the controversies that have raged over them, I think this objection *prima facie* fails, even if it is hard to explain why. Those inquirers engaged in explicitly fictive thinking are not aiming to write fantasy novels. Rather, the aim of fictive thinking is expedience, in terms of both practical activity and in terms of the organization of thought and the direction of inquiry. Therefore, fictions have a connection with both the use of theoretical reason and practical reason. This means that they engage our interest in both the true and the good, though at a certain remove from the immediacy of either belief or action.

5.3. "Why not 'acceptance' or 'supposition' instead of fiction?"

Any propositional attitude that is a "hypothesis" in Vaihinger's sense, something that affirms something about reality (such as knowledge, opinion) or is intended as a tentative or hopeful precursor of an affirmation (such as acceptance, supposition), is subject to a parallel of conciliationism. If disputants have apparently reasonable disagreements in their opinions or suppositions, or over which claims to accept, this likewise acts as a second order piece of evidence undercutting whatever evidence supported the differing opinions or suppositions or acceptances. That Columbo and Poirot have different opinions or suppositions (rather than beliefs) about whodunit is good reason for both to revise their opinions and suppositions about the matter. Acceptances are more akin to fictions in that they are objects of volition (Cohen 1992, 22). Yet whilst it seems justifiable to adopt one of a number of different acceptances in adverse epistemic conditions, doing so is not justifiable as epistemic conditions improve; acceptances are therefore sensitive to evidence in a way that fictions are not. For instance, it seems justifiable for one physicist to accept one version of string theory, and for another physicist to accept some other version of string theory, as temporary propositions. But it is not justifiable for a contemporary physicist to accept in this way Newtonian physics, whereas it remains very expedient to take it up as a fiction in many contexts.

5.4. "In terms of ecological expediency, are there not also circumstances in which it is essential that everyone is guided by the same fictions? e.g. that we all employ the fiction of free will."

When speaking about expediency, I have been speaking principally about the expediency regarding the advancement of intellectual understanding. The example of juridical punishment is one in which we aim at something more concretely practical, where it seems true that we must all converge on the same fiction. But since fictions are non-rivalrous and are justified by expediency, there is no problem with our employing one fiction at one time or context and another fiction at another time or context depending on the purpose in hand. For instance, a judge *qua* judge will act and think as if the convict was free in committing their crime and merits a certain punishment. Such a fiction is part and parcel of the role of judge and the practice of juridical punishment. But the same judge might be quite committed to determinism$_f$ when as a prudential agent he has to decide which part of the city to reside in.

5.5. "When I say 'I believe God exists' or 'I believe abortion is murder' I am stating my beliefs. I do not mean 'I think and act as if God exists,' and I refuse to mean this. Your suggestion is unacceptable regarding matters such as religion and morality."

One response to this objection is to aver that many of one's epistemic peers do not believe that God exists or that abortion is murder, and that the objector ought to fictionalize their religious and moral commitments. But two other responses can be made, each of which avoids the demand for fictionalization. A first is that religious and moral beliefs are not founded on evidence at all. Whilst such a response might raise questions about whether religious and moral beliefs ought to be held at all, it indicates that disagreements about religion and morality are not apparently reasonable. If the beliefs of A are not based on the reasoned consideration of a body of evidence, and B is aware of this, then the fact that A endorses certain beliefs provides no undercutting defeater of B's beliefs. Second, one might question whether the commitments of religion and morality are best characterized in terms of belief. For example, many philosophers who are by no

means hostile to religious faith have characterized it as being something other than a species of belief. For Kant, faith is a practical mode of conviction distinct from theoretical knowledge and opinion (Wood 1970, 17-25). Robert Audi describes faith as a cognitive attitude separate from belief, and one which is "epistemically less at risk, in the sense that it is less easily defeated, than rational belief" (Audi 1991, 219). Going further, Schellenberg claims that belief and faith are incompatible (Schellenberg 2005, 127-166), whilst for Schleiermacher faith is a feeling rather than a cognition (Schleiermacher 1893). Similar claims can be made about some (but by no means all) metaethical views; familiar kinds of non-cognitivism would plausibly avoid the need for conciliation and fictionalization, as would (more arguably) certain varieties of constructivism, intuitionism, and moral sentimentalism.

Acknowledgments

Thanks to participants at the Louisiana State University Graduate Philosophy Conference 2018 for their comments on an earlier version of this paper. Thanks also to the reviewers and editors for their time.

References

AUDI, R. (1991): Faith, Belief, and Rationality. *Philosophical Perspectives* 5, 213-239.

CARTWRIGHT, N. (1983): When Explanation Leads to Inference. *Philosophical Topics* 13(1), 111-121.

CHRISTENSEN, D. (2009): Disagreement as Evidence: The Epistemology of Controversy. *Philosophy Compass* 4(5), 756-767.

COHEN, L. J. (1992): *An Essay on Belief and Acceptance*. Oxford: Oxford University Press.

CRAIG, E. (1991): *Knowledge and the State of Nature*. Oxford: Oxford University Press.

DESCARTES, R. (1996): *Meditations on First Philosophy*. (J. Cottingham, Ed.). Cambridge: Cambridge University Press.

FAHRBACH, L. (2011): Theory Change and Degrees of Success. *Philosophy of Science* 78(5), 1283-1292.

FEYERABEND, P. (1975): *Against Method: Outline of an Anarchist Theory of Knowledge*. New York: Verso.

FORBERG, F. (2010): Development of the Concept of Religion. In: Y. Estes & C. Bowman (Eds.): *J.G. Fichte and the Atheism Dispute* (31-48). Farnham: Ashgate.

HACKING, I. (1982): Experimentation and Scientific Realism. *Philosophical Topics* 13(1), 71-87.

HARDIN, C., & ROSENBERG, A. (1982): In Defense of Convergent Realism. *Philosophy of Science* 49(1), 604-615.

HAYEK, F. (1955): Degrees of Explanation. *The British Journal of the Philosophy of Science* 6(23), 209-225.

HOBBES, T. (1998): *Leviathan*. Oxford: Oxford University Press.

HUME, D. (2007): *A Treatise of Human Nature, Volume 1*. (D. Norton & M. Norton, Eds.). Oxford: Oxford University Press.

KELLY, T. (2005): The Epistemic Significance of Disagreement. In: T. Gendler & J. Hawthorne (Eds.), *Oxford Studies in Epistemology, Volume 1*. Oxford: Oxford University Press.

KING, N. (2012): Disagreement: The Skeptical Arguments from Peerhood and Symmetry. In: D. Machuca (ed.): *Disagreement and Skepticism*. Abingdon: Routledge.

LAUDAN, L. (1981): A Confutation of Convergent Realism. *Philosophy of Science* 48(1), 19-49.

PLANTINGA, A. (1995): Pluralism: A Defense of Religious Exclusivism. In: *The Rationality of Belief and the Plurality of Faith*. Ithaca: Cornell University Press.

PSILLOS, S. (1999): *Scientific Realism: How Science Tracks Truth*. New York: Routledge.

RAWLS, J. (1999): *A Theory of Justice* (Revised Ed). Cambridge, MA: Harvard University Press.

SCHELLENBERG, J. L. (2005): *Prolegomena to a Philosophy of Religion*. Ithaca: Cornell University Press.

SCHLEIERMACHER, F. (1893): *On Religion: Speeches to its Cultured Despisers* (Third Ed). London: Kegan Paul, Trench, Trübner & Co.

VAIHINGER, H. (1935): *The Philosophy of "As If": A System of the Theoretical, Practical and Religious Fictions of Mankind* (Second Ed). New York: Harcourt, Grace, and Company.

WILLIAMS, B. (2002): *Truth and Thruthfulness: An Essay in Genealogy*. Princeton, NJ: Princeton University Press.

WOOD, A. (1970): *Kant's Moral Religion*. Ithaca: Cornell University Press.

WORRALL, J. (1989): Structural Realism: The Best of Both Worlds? *Dialectica* 43(1), 99-124.

Organon F 25 (4) 2018: 473-494

Should a Causal Theory of Reference Borrowing Be a Descriptive-Causal Theory?

LUIS FERNÁNDEZ MORENO[1]

ABSTRACT: In a reference theory a distinction can be made between a theory of reference fixing and a theory of reference borrowing. M. Devitt and K. Sterelny, and especially the former, have been relevant figures in the present debate on reference theories. They have supported a descriptive-causal theory of reference fixing for proper names and natural kind terms, but they have held a purely causal theory of their reference borrowing. Once I have put forward the main elements of Devitt's and Sterelny's theory of reference fixing I will focus on their reference borrowing theory. In this regard I will examine some of the differences between Devitt's and Sterelny's causal theory of reference borrowing and Putnam's thesis of the division of linguistic labor concerning natural kind terms. After taking into consideration the views of some causal theorists who have not rejected or have even explicitly admitted that there are descriptive requirements in a reference borrowing theory for proper names and natural kind terms, I will allege that a causal theory of reference borrowing for competent speakers should not be a purely causal theory, but a descriptive-causal theory, where the minimum descriptive component is some general categorial term that is true or approximately true of the referent of the term.

KEYWORDS: Categorial term – descriptive-causal theory – natural kind term – proper name – reference borrowing – reference fixing.

[1] Received: 15 June 2018 / Accepted: 24 August 2018

✉ Luis Fernández Moreno

Department of Logic and Theoretical Philosophy, Faculty of Philosophy
Complutense University of Madrid, 28040 Madrid, Spain

e-mail: luis.fernandez@filos.ucm.es

1. Devitt's and Sterelny's theory of reference fixing and reference borrowing

A theory of reference provides the answer to the question of how expressions get connected (that is, refer) to an entity or to certain entities. However, in such a theory a distinction can be made between a theory of reference *fixing*, which explains how the referent of a term is initially determined, i.e., by the speaker(s) who introduced it, and a theory of reference borrowing, which explains how the reference of the term is determined for the rest of the speakers.

In (1999), a classic book in contemporary Philosophy of Language, which I will treat as the backbone of my considerations, Devitt and Sterelny support a purely causal theory of reference borrowing for proper names and natural kind terms, but they claim that the theory of the ostensive reference fixing[2] for both sorts of expressions[3] has to include *descriptive* components and thus be a *descriptive-causal* theory. Since I will allege that some of those descriptive elements should also be involved in a theory of reference borrowing, I will first pay attention to the main constituents of Devitt's and Sterelny's theory of reference fixing, and as they deal with proper names before natural kind terms, I will begin by taking into consideration their reference fixing theory for proper names.

At the basis of the quandary that has led Devitt and Sterelny to sustain that a theory of reference fixing for *proper names* must contain descriptive

---

[2] Since Devitt's and Sterelny's theory of reference fixing for proper names and natural kind term focusses on their ostensive introduction and thus on their ostensive reference fixing, this is the only sort of reference fixing theory I will be taking into consideration.

[3] In regard to (paradigm) proper names Devitt and Sterelny use the term "designation" and concerning general terms – and hence natural kind terms – the term "application". Instead of these expressions I will generally use the term "reference", which they employ for "the genus of which all referential relationships – for example, application, designation, denotation – are species" (Devitt & Sterelny 1999, 312). On the other hand, in the case of proper names the vehicles of reference for Devitt and Sterelny are name tokens (or uses of names); however, I will often simply speak of proper names or names. It is to be assumed that a similar consideration would apply to natural kind terms, although concerning them I will also speak of that sort of terms and not of their tokens (or uses), and with respect to this type of terms I will use indistinctly the notions of reference and extension.

components underlies the ambiguity of the ostension to the object involved in that reference fixing, but by means of the name we intend to refer to the object *qua whole object*. The *disambiguation* of the reference requires, according to Devitt and Sterelny, that the introducer of the name *conceptualize* the object by means of "some general categorial term" (Devitt & Sterelny 1999, 80), in such a way that if he were very wrong about it, the name would *lack* reference. Thus, they assert that the theory of reference fixing for proper names "must be a 'descriptive-causal' theory: a name is associated, consciously or unconsciously, with a description in a grounding" (Devitt & Sterelny 1999, 80).[4] Regarding this, it should be assumed that such description is a demonstrative one that contains the general categorial term in question.

By *categorial term* Devitt and Sterelny seem to understand sortal terms in the broad sense,[5] that is, terms that convey a criterion of *identity*. When they use the expression "general categorial term" they allude to highly

---

[4] In order to allude to an initial baptism (initial introduction) of a term in the sense of Kripke's (Kripke 1980), Devitt and Sterelny use the term "dubbing", and they call "dubber" the (initial) introducer of the term. They understand by "grounding" "a perception […] of an object that begins a reference determination causal chain for a term" (Devitt & Sterelny 1999, 310). The notions of grounding and dubbing are related, and therefore those of grounder and dubber, since a dubbing of a term is the initial grounding of the term – Devitt and Sterelny claim that there are usually multiple groundings of a term (see note 14 below). Since the notion of grounding is more general than the one of dubbing, in the rest of the paper I will mainly use the former, but I will speak of dubbing or dubbers when it is required to emphasize that I mean the initial grounding or grounders.

[5] Sortal terms in the strict sense provide criteria for individuation and for identity concerning the entities to which they apply; thus count terms are sortal terms in the strict sense. On the contrary, mass terms have no criterion for individuation governing their application, although they do have one of *identity*. They are not sortal terms in the strict sense, but some authors occasionally use the notion of sortal term in a broad sense (see, e.g., Hale & Wright 1997, 685), according to which the distinguishing feature of sortal terms should be to convey a *criterion of identity*, a feature shared by count terms and mass terms. I will denominate the terms possessing this feature *categorial terms*, following some of the suggestions in Devitt & Sterelny (1999, 80 and 90; see also Devitt 1981, 63 f.). These terms can be simple or complex, and they form indefinite descriptions with an indefinite article, and in the case of categorial mass terms, with the further aid of certain classifier phrases.

general terms of that type.[6] On this matter it is worth mentioning that they give the terms "animal" or "material object" (Devitt & Sterelny 1999, 80) as examples of the general categorial terms involved in the grounding of a name for a cat; however the term "material object" will not generally fulfill the role of disambiguating the reference.

The reference fixing theory put forward by Devitt and Sterelny in (1999) concerning *natural kind terms* bears similarities with their theory about proper names, but there are some differences, especially that the grounding of a natural kind term involves a perceived *sample* of objects of the kind and that it includes "an ostensive component and a 'nature' component" (Devitt & Sterelny 1999, 88).

The extension (reference) of a natural kind term will contain all the samples of the *same kind*, i.e., those that share the same *underlying structure or nature* as the ostensively given sample in that grounding. Devitt and Sterelny claim that the reference fixing (grounding) of natural kind terms involves us in the *qua*-problem, whose source is that "the term is applied to the sample […] *qua* member of a natural kind but also *qua* member of one particular kind" (Devitt & Sterelny 1999, 91). Since two parts can be distinguished in this *qua*-problem, these authors allege that in order to sort out this problem, providing the required disambiguation of the entities given in the grounding, *two descriptive components* are rendered necessary. However, to our aim, what is especially relevant is the first part of the *qua*-problem, whose solution requires that "the grounder of a natural kind term associates, consciously or unconsciously, with that term […] some description that in effect classifies the term as a natural kind term" (Devitt & Sterelny 1999, 92).[7] The aim of this descriptive component is to establish that the term to be introduced is a *natural kind term*.[8] Although

---

[6] In this paper I will use the expression "general categorial term" in this sense. Categorial terms are general terms, but by the use of that expression I will refer to highly general categorial terms.

[7] In the proposed solution for the *qua*-problem concerning proper names and natural kind terms Devitt and Sterelny assert that the association of descriptions with such expressions takes place "consciously or unconsciously". By this I understand that the association in question can be implicit or explicit. I will assume this view.

[8] The solution to the second part of the *qua*-problem requires that the grounder associates with the natural kind term "some descriptions that determine which nature of the

these authors are not very definite in this regard, it is to be assumed – by parallelism with the *qua*-problem for proper names – that in the formulation of the descriptions that classify a term as a natural kind term and in which it is appealed to entities given ostensively it is necessary to resort, implicit or explicitly, to some *general categorial term* that conceptualize those entities as members of a natural kind.[9]

As we have said, Devitt and Sterelny hold a *purely causal* theory of reference borrowing. The basic idea of their theory for (paradigm) proper names is that by virtue of the causal-perceptual link between a name and an object, the grounder, and other speakers at the grounding, acquire the *ability* to use the name to refer to the object. Those speakers will use the name in conversation with others and the latter will acquire the name and borrow its reference from the former speakers – the *lenders* – by acquiring from them that sort of ability on the basis of the perception of the use that the lenders make of the name; thus the acquisition of the *borrowers*' ability to refer to an object by a name also involves a causal process. In this way those latter speakers – by the exercise of that sort of semantic ability – will refer to the object in virtue of causal chains that link the object with uses of the name caused by the ability acquired from the lenders. So the name will be transmitted through the linguistic community at the same time as the abilities to use the name to refer to the object are passed on, and as new links are added to the causal chains involving the uses of the name, which constitute a causal network – the latter usually being multiply grounded in the object (see note 14 below). However, the properties that borrowers

---

sample is relevant to the reference of the term." (Devitt & Sterelny 1999, 92). The aim of this second sort of descriptive component is to determine *which* of the natural kinds the sample belongs to will be the referent of the term – the sample will usually belong to several natural kinds of different generality, like gold, metal, element, etc. On this matter, Devitt and Sterelny claim that it is necessary to resort to descriptions of certain *macroscopic properties* of such objects, in particular, of observable properties and causal powers macroscopically discernible – see Devitt & Sterelny (1999, 92) and Sterelny (1983) –, since the relevant underlying structure will be the one responsible for such properties.

[9] Thomasson (2007, chapter 2), holds the view that the reference fixing of proper names and general terms, like natural kind terms, requires that those terms be associated with categorial terms, which contribute to disambiguate the intended reference. This author also replies to criticisms against the indispensability of that sort of descriptive requirement.

could associate with a name *do not determine* the reference of the name as they use it, since they need not be possessed by the referent of the name.

Devitt's and Sterelny's theory of reference borrowing, first put forward for proper names, is similar to the one concerning natural kind terms. Likewise, as in the case of proper names, the properties that borrowers could associate with a natural kind term do not determine the reference of the term as they use it, since it is *not necessary* that those properties be possessed by the entities of the kind.[10]

Devitt and Sterelny have made several explicit assertions concerning the absence of descriptive requirements in the reference borrowing. For example, they assert that "the pure-causal theory of reference borrowing does not require borrowers to associate with a term any [true] description of its referent. This […] [is] appropriate for names and natural kind terms" (Devitt & Sterelny 1999, 97). They also claim that "borrowers do not have to associate the *correct categorial term*" (Devitt & Sterelny 1999, 80; emphasis added). In a later writing Devitt asserts that "the theory of reference borrowing shows how a person can be linguistically competent with a word despite being *largely* ignorant, or even wrong, about its referent. People can be competent with the name 'Catiline' despite knowing *very little* about Catiline" (Devitt 2006, 139; emphases added).

In (1999) Devitt and Sterelny claim that the competence with a term consists in the ability acquired in a grounding or reference borrowing to use a term to refer to an object or to samples of objects. According to them *all* reference borrowers of a term (proper name or natural kind term) are competent with the term, but even accepting this claim for the sake of the argument – see section 3 –,[11] the question arises whether the borrowers'

[10] Kripke's theory of reference borrowing for proper names and natural kind terms is also purely causal, since the reference of a term as used by borrowers is exclusively determined by their membership in a causal chain *independently* of the descriptions or properties they could associate with the term, since these do not play any role in such reference determination (see Kripke 1980). This does not exclude that the borrowing includes an intentional component, as it also happens in Devitt's theory (see section 2).

[11] According to Devitt and Sterelny the reference borrowing with respect to a term entails competence concerning the term. Since in section 3 I put that claim into question I will sometimes use the expression "competent borrowers". This expression can seem redundant in the case of Devitt's and Sterelny's theory but it leaves the option open that there are borrowers who are not competent, which is my view.

linguistic competence with a word is compatible with great ignorance or error about its referent, and in case that for this competence it is required to know "very little" about the referent, what is the descriptive component that competent borrowers have to associate with the term. My proposal will be that the minimum component required of competent borrowers will be some *general categorial term* (or equivalently the corresponding property) which is true or approximately true of the referred entity.[12]

2. Devitt and Sterelny on Putnam's division of linguistic labor

Although Devitt and Sterelny claim that competent speakers acquire an ability to refer to members of a natural kind in a grounding or reference borrowing, it is plausible that the speakers that acquire a better or rather a more reliable ability to refer to the members of a natural kind are the *experts*, in Putnam's sense, one of the main notions of his (hypo)thesis of the division of linguistic labor concerning natural kind terms.[13] Since Devitt and Sterelny, and especially the former, have made some remarks on

[12] Jutronić in (2008) has held a similar although stronger view; she claims that "reference borrowing involves the borrowers having to associate *the correct* categorial term and have some *true* beliefs about the referent in the guise of some associated description" (2008, 358 emphases added). However, my argumentation is different and independent from hers. In any case, I would like to make two comments. Firstly, there is usually more than one categorial term associated with a proper name or a natural kind term that could be considered as true or correct of its referent. Secondly, I would introduce the caveat "approximately true" that I have taken from Putnam (see section 3) – and in the same way "approximately true" beliefs, also used by Putnam. The aim of this nuance is to block the arguments from the ignorance-and-error type that could be put forward concerning borrowers, although Putnam does not present his proposal in this framework.

[13] On this matter two comments are adequate. Firstly, after Putnam (1975c), like in (1988), Putnam does not talk any longer of the "hypothesis of the universality of the division of linguistic labor", but simply of the *division of linguistic labor*, assumed as a thesis he subscribes (see Putnam 1988, 22), and so I will consider it in the following. Secondly, Putnam sometimes speaks, and Devitt and Sterelny always do (see below), of the "linguistic division of labor" (see, e.g., Putnam 1975c, 274 and 1988, 25) instead of the "division of linguistic labor". Putnam uses both expressions interchangeably (see Putnam 1988, 22, 25 and 37).

Putnam's central notions involved in that claim, it seems appropriate to present some of Putnam's views before taking those comments into account.

As is well-known, in his theory concerning how the reference or extension of natural kind terms is determined (see Putnam 1975c) Putnam underlines two contributions, to which he alludes as the *contribution of the environment* and the *contribution of the society* – on these contributions see Putnam (1975c, 227-234, 245, 265 and 271), as well as (1988, chapter 2). On the one hand, the extension of a natural kind term depends on how our environment or our world is since it is determined by underlying properties of the members of the kind belonging to our world. On the other hand, the discovery of the underlying properties is a matter of scientific research and those who carry it out or, in a more general way, who are able to apply reliable *tests* to distinguish members of a natural kind from entities not belonging to it – those are called "experts" by Putnam – are more knowledgeable than the average speaker concerning the membership conditions into a natural kind and hence into the extension of the corresponding natural kind term. There is in this regard, in Putnam's words, a *division of linguistic labor*, in such a way that the average speakers or non-experts rely on experts and are willing to *defer* to experts concerning the determination of the reference of natural kind terms as they use them. Thus, an entity falls into the extension of a natural kind term used by the average speaker if it falls into the extension of the term as used by the experts and the average speaker is linked by relationships of cooperation or rather of links of *deference* with the experts in question.

Concerning the notion of expert it is relevant to point out that, although members of the relevant scientific community are experts *par excellence*, the group of experts has to be conceived in a *broad* sense, since the experts are those members of our linguistic community that are able to apply the mentioned tests, but those tests do not need to include the explicit description of the underlying properties of the members of the kind – in case they are known. In this way, for instance, concerning the word "gold" Putnam considers as experts on gold not only atomic physicists and chemists but also metallurgists, miners and jewellers (see Putnam 1996, XVI).

The main notions of the thesis of the division of linguistic labor are those of expert and deference. Regarding this, Devitt and Sterelny claim that although the dubbers of a natural kind term may be experts in Putnam's sense,

they need not be so, although at least some of the experts will be later *grounders* of the term.[14] They also assert in (1999) that "Putnam brings out the significance of reference borrowing by talking of 'the linguistic division of labour'" (Devitt & Sterelny 1999, 88), and Putnam seems to agree with that claim, since in a paper on Devitt's views Putnam identifies the division of linguistic labor with the reference borrowing (Putnam 2001, 498).

On this matter it is relevant to point out that in a passage of Devitt (2006), he comes to *identify* reference borrowing with deference (Devitt 2006, 138), but later he *distinguishes* them for two reasons:

> (1) If *x* borrows the reference of a term from *y*, then that is an act at the time of receiving *y*'s communication. In contrast, if *x* defers to *y*'s use of a term, then that suggests an act at the time of *x*'s using the term herself to communicate. (2) Furthermore, talk of deference invites a confusion between *epistemic* deference to experts *when seeking knowledge*, which we should all be in favour of, and *semantic* deference to experts *when referring*, which causal theorists oppose. (Devitt 2015, 116-117)

When characterizing the notion of deference in the first part of this passage, Devitt does not allude to Putnam. However, when Devitt speaks in the second part of the passage about deference he is taking into consideration Putnam's notion of expert. Therefore, it is to be assumed that what Devitt asserts on deference in the first part of the passage would also concern Putnam's theory.

---

[14] As already said in note 4, Devitt and Sterelny claim that there are usually multiple groundings of a term. They put forward the thesis of multiple grounding concerning proper names as follows: "many uses of a name are relevantly similar to a dubbing […] [since] they involve the application of the name to the object in a direct perceptual confrontation with it […]. Such uses of a name ground it in its bearer just as effectively as does a dubbing. As a result it becomes multiply grounded" (Devitt & Sterelny 1999, 75). And these authors hold a similar thesis concerning natural kind terms: "Multiple grounding is important with natural kind terms, as it was with names […] [A] natural kind term is grounded just as effectively [as in the dubbing] by subsequent groundings" (Devitt & Sterelny 1999, 89-90.). Devitt and Sterelny proposed the thesis of multiple grounding mainly to explain the changes of reference that proper names and natural kind terms may experience, but I cannot enter into this matter here.

However, Putnam never held, what seems to be implied by Devitt's claim (1), that the speaker defers to experts *whenever using* a term (to communicate). The deference to experts in Putnam's sense involves the speaker's intention to defer to experts, but in Putnam's theory there is no clear indication concerning when that deference takes place. It can happen at the time when the speaker acquires or uses a term for the first time but it can also concern *further* speaker's uses of the term, although not necessarily all of them.

It is worth mentioning that Devitt distinguishes between the initial borrowing and the later use of the borrowed term; although both are intentional acts,[15] in Devitt's causal theory of reference borrowing, at least after his version in (Devitt 2008), the resort to the notion of deference does not play any role therein, since "according to the causal theory, the later use *need not involve any intention to defer to the earlier borrowing*" (Devitt 2008, 362); furthermore, none of the uses of a term by the borrower require to be accompanied by the intention to defer, and the borrower "need not defer to the lender" (Devitt 2015, 116). Thus, in Devitt's theory the borrowing of a term does not need to involve deference. In my view, Devitt's assertion (1) in the quoted passage from Devitt (2015, 116-117) can only be justified in the sense that reference borrowing is not the same as deference; therefore the equivalence of reference borrowing with deference and with the linguistic division of labor should be rejected. The rejection of the relevance of (semantic) deference is alleged more clearly in the second part of the passage, i.e., in assertion (2).

However, in this assertion Devitt is *assuming* his version of the causal theory which in this regard agrees with Kripke's theory, according to which deference to experts in Putnam's sense – the core of the division of linguistic labor – does not play any role in the determination of reference.[16] On

---

[15] Kripke's theory also embodies an intentional component, since he claims that for a borrower, or rather for the use of a name by a borrower, to be a link of the causal chain involving a name – and the same will hold concerning natural kind terms – it is required that, when he learns the name, he *intends* to use it with the *same reference* as it was used by the speaker from whom he learnt it (see Kripke 1980).

[16] It is noteworthy that Kripke does not accept the notion of expert in Putnam's sense; for Kripke the only experts concerning the reference of terms are those speakers who have introduced the terms in an initial baptism (see Kripke 1986).

this subject, in (1999) Devitt and Sterelny allude to the grounders as "experts", between inverted commas, for example, in assertions of the sort "the 'experts' who fix the reference" (Devitt & Sterelny 1999, 97), and distinguish them from experts in Putnam's sense – with no quotation marks – alleging that "[the] grounders may be experts […], but it is not essential that they be" (Devitt & Sterelny 1999, 89). However, according to what was said above, at least in works from Devitt (2008) Putnam's thesis of the division of linguistic labor, which involves the (semantic) deference to experts, does not play any *role* in Devitt's reference theory, but it would be a mistake if the reason for that should be the confusion alleged by Devitt in assertion (2). In my view there is no such confusion; indeed, the experts are those speakers who *know* more than the average speaker concerning natural kinds, since they have better tests to distinguish members of a natural kind from entities not belonging to it and therefore better criteria for the *reference* of natural kind terms than the average speaker. The supposed confusion arises from *assuming* the sort of causal theory of reference Devitt sustains, but in that regard Devitt is begging the question.

It is worth mentioning though, that in works in which Devitt and Sterelny still accepted the relevance for a reference theory of Putnam's thesis of the division of linguistic labor, after asserting that, as mentioned above, "Putnam brings out the significance of reference borrowing by talking of 'the linguistic division of labor'" (Devitt & Sterelny1999, 88; see also Devitt 2006, 138 and note 161), Devitt and Sterelny describe a scenario in which an *apprentice jeweller* learns the term "platinum" from another speaker. But since an apprentice jeweller is an apprentice expert,[17] it may be supposed that he learns the term "platinum" from an *expert* who shows him a sample of platinum uttering, as those authors say, the words "That is platinum", but in that scenario there is no allusion at all to deference in Putnam's sense. Devitt and Sterelny assert:

> Consider the case of an apprentice jeweller learning the term 'platinum'. A sample of platinum is pointed out to him with the words, 'That is platinum'. He gains an ability to use the term to refer to platinum, an ability grounded in the metal by this introduction. His later uses of the

[17] As already said Putnam claims that among the experts on gold – and the same will apply to platinum – are jewellers (see Putnam 1996, XVI).

> term, exercising that ability, will refer to the metal in virtue of their causal link to it. (Devitt & Sterelny1999, 89)

This passage gives rise to several remarks. Firstly, the question arises whether the apprentice jeweller gains the ability to refer to platinum *exclusively* by perceiving a sample of platinum accompanied by the utterance "This is platinum" and hence whether the apprentice expert's later uses of the term will designate samples of platinum solely by virtue of his causal link to that sample. Although the answer to these questions concerning the said apprentice jeweller would be affirmative if Devitt's and Sterelny's causal theory of reference borrowing is accepted (however, see the second paragraph below), the answer should be negative if the thesis of the *division of linguistic labor* is assumed, as Devitt and Sterelny did in (1999). The answer is negative because to the extent that the apprentice expert is learning the term "platinum" from an expert, his ability to refer to platinum will be at least partly determined by the beliefs or knowledge about platinum – and thus, by *descriptions* – that he learns from the expert, which will involve tests to identify samples of platinum and distinguish them from samples of other substances; this is a condition to be an expert on platinum from whom the apprentice expert is acquiring the ability to refer to platinum, and to whom the apprentice is *deferring*.

Secondly, Devitt's and Sterelny's proposal regarding the first sort of the *qua*-problem involved in the reference fixing or *grounding* of natural kind terms is applicable to the scenario they present in the quoted passage. The use of a term by an expert – "That is platinum" – is another grounding ("introduction" is said in the passage) of the term "platinum" and in this regard let us bear in mind, as mentioned above, that there are usually multiple groundings of a term. As already said, according to Devitt and Sterelny in the grounding of a natural kind term in which it is appealed to entities given ostensively it is necessary to resort, implicitly or explicitly, to some *general categorial term* that conceptualize those entities as members of a natural kind.

Thirdly, that thesis about grounding has consequences on borrowing. Regarding this, if we focus on the act of pointing to the sample of platinum in question uttering in that context the words "That is platinum", it can be claimed that, given the ambiguity of the ostension, the demonstrative "that" must be supplemented with some general categorial term that disambiguates

the particular sample concerned; in this example the term could be, e.g., "metal": "That (sample of) metal is (a sample of) platinum". And since in the said context the borrower is learning the term "platinum" and borrowing the reference of the term, he will associate with the term "platinum" a general categorial term, like the term "metal".

Thus, even leaving aside the more specific beliefs or knowledge of the expert from whom the apprentice jeweller is borrowing the term "platinum", it can be argued from a more general level that some general categorial term is required for those cases of reference borrowing which involve the ostension to a sample of the term's referent. Therefore, the theory of reference borrowing on natural kind terms, at least in the case contemplated by Devitt and Sterelny in (1999) regarding the apprentice expert, and some features of this case can be generalized, should be *descriptive-causal* and not purely causal.

3. A moderate epistemic view of the reference borrowing for competent speakers

Although Devitt and Sterelny do not pay attention to this fact, there are several *causal theorists* who have not rejected, or have even explicitly admitted, the thesis that there are descriptive requirements in a theory of reference borrowing. I will take into consideration two of them, K. Donnellan and H. Putnam.

Donnellan, one of the main advocates of the *historical-causal theory*, does not question the necessity to incorporate *descriptive* components in a borrowing reference theory for proper names, whose claims should extend likewise to that sort of theory concerning natural kind terms; for this reason I will sometimes speak simply of "terms".

In Donnellan's most famous paper devoted to criticizing the description theory of reference on proper names, i.e., Donnellan (1972), he does not dispute the claim that it may be a necessary condition – although not a sufficient one – for an entity to be the referent of a term as used by the *borrowers* that such an entity satisfy some description that they associate with the term. However, he considers that it is too strong a requisite to demand that this description has to be an identifying description, i.e. a description sufficiently specific to uniquely identify one individual (Donnellan

1972, 366-367). In this regard, he does not find the claim objectionable "that our use of the name ['Aristotle'] is such that being a human being or not living in modern times, etc. are *necessary* for being the referent of the name" (Donnellan 1972, 367).

In this passage Donnellan does not reject the thesis that there are descriptions or general terms associated with a term which may be considered necessary for an entity to be the referent of the term; among them are those that express the *type* of entity referred to as well as other general properties of the referent. Nonetheless, since in the case of different individuals the second class of properties can be very different, the most comprehensive unquestioned property is that of being a type of individual or entity, in the example of the name "Aristotle" the property of being a human being, where the term "human being" is, of course, a general categorial term.

Let us take Putnam into consideration, an advocate of the *causal-social theory* – he calls his view of reference a "causal/social outlook" (Putnam 1975d, 281). This author holds more definitely the requirement that the borrower must associate some descriptive components with the borrowed term. Although Putnam has not proposed a theory of proper names, he claims that "unless one has some *beliefs* about the bearer of the name that are *true or approximately true*, then it is at best idle to consider that the name refers to that bearer in one's idiolect" (Putnam 1975b, 203; emphases added). Concerning this, he gives the following example: "I do not see much point, for example, in saying that someone is referring to Quine when he uses the name 'Quine' if he thinks that 'Quine' was a Roman emperor, and that is all he 'knows' about Quine" (ibid.). However, this is compatible with the claim that the speaker associates with the term "Quine" some "minimal linguistic information […], namely that it is a person's name" (Putnam 1975b, 201). Thus Putnam is not questioning that the speaker associates with the name "Quine" the general categorial term "person" or, what is relevant for our considerations on Donnellan's example and other subsequent ones, the term "human" or "human being" – or the corresponding properties.

I agree with a similar claim to the one put forward by Putnam in the passage from (Putnam 1975b, 203), according to which some of the descriptions or properties that users of a term and especially *borrowers* associate with it must be true or *approximately true* of the entity that constitutes its referent for those speakers to refer to that entity. In this respect, it is

relevant to make at least two remarks. Firstly, in that passage Putnam is *not* speaking about competence, but only about users of terms (to refer), who include reference borrowers; thus, from that passage nothing is derivable about Putnam's view on the conditions for a speaker to be competent regarding a proper name.[18] However, since according to Devitt and Sterelny reference borrowing entails competence, the claim by Putnam (in 1975b, 2003) could be extended within the framework of Devitt's and Sterelny's theory to that group of competent speakers, the "competent borrowers" (see note 11). Secondly, the term "approximately true" is not in contexts of this sort susceptible of a precise analysis, but although Putnam does not say so, the aim of introducing the nuance "approximately true" is in my view, as already said in note 12, to block the arguments from the ignorance-and-error type. Nevertheless, if someone asked me for an example of a property approximately true of an entity, I would put forward the following example. Let us assume that, although I and the people around me do not know it, my friend Richard is in fact a very sophisticated robot, not a human being, but with the external behaviour, all of the external features and some of the internal ones, even emotional feelings, characteristic of a human being. The property of being human would not be true of Richard, but would be approximately true of him, since he shares many properties with human beings. To those considerations underlies the view that the borrowers cannot be *completely ignorant* or *wrong* about the properties of the entity they refer to; e.g., if a borrower, who had learnt the name "Richard" in a purely causal way, would associate with the name "Richard" the property of being a river, a mountain, a building,… or only properties that do not apply at all to Richard, we could allege that we lack any justification to consider that by means of the name the borrower is referring to such an entity, or in Putnam's words, "it is at best idle to consider that the name refers to that bearer in one's idiolect." Of course, in this field, like in most fields in philosophy, there are no arguments that definitively decide a question, but only plausibility claims, and I consider Putnam's view plausible. In any case, we have already quoted a passage from Devitt (2006, 139), where he

[18] Nevertheless, from his assertion in Putnam (1975b, 201) it could be alleged that according to Putnam a competent speaker regarding the name "Quine" is to associate with it the general categorial term "person".

asserts about that sort of competent speakers that according to him reference borrowers are, that "[borrowers can be] largely ignorant, or even wrong, about its referent [the referent of a word]", but that is not the same as being *completely* ignorant or wrong about it, and he concedes in the same passage that borrowers can know "very little" about the referent. My answer to the question of what that "very little" can consist of is that the reference borrowers regarding a term have to associate with it at least the property of being the type of entity that the referent is, which is expressed by some general categorial term and the latter has to be true or approximately true of the referent.

However, in his theory of natural kind terms Putnam speaks more explicitly of *competence*, and according to Putnam's view concerning this sort of terms, all competent speakers will have to associate with a term, implicitly or explicitly, the syntactic markers, the semantic markers and the stereotype of the term (see Putnam 1975c). The most relevant of these factors for this paper are the last two, although Putnam claims that "in the extreme case, the stereotype may be *just* the [semantic] marker: the stereotype of molybdenum might be *just* that molybdenum is a *metal*" (Putnam1975c, 230), where the property of being a metal indicates the type of entity that molybdenum is. In the cases in which the stereotype is different from the semantic markers, the main feature to distinguish the second from the first is that the semantic markers are "category-indicators of high centrality" (Putnam 1975c, 268) and hardly revisable, although semantic markers as well as the properties included in the stereotype are not analytically associated with the natural kind term in question. However, according to Putnam these properties must be associated with the term for the speaker to be competent concerning that term. And this claim also applies to the reference *borrowers* insofar as they are competent speakers. In his (1975c) Putnam mainly details the syntactic markers, the semantic markers and the stereotype concerning the term "water". In this case the stereotype includes many properties, since the average competent speaker associates many of them with the term "water" – colorless, transparent, tasteless, thirst-quenching, etc. (Putnam 1975c, 269) –, but in other cases – and this happens concerning many natural kind terms, and not only in extreme cases – the stereotype will coincide with the semantic markers, and since we are interested in the question of whether there are descriptive components involved in the reference borrowing, in this case of natural kind terms, the

answer will be affirmative if at least the properties contained in the *semantic markers* (or at any rate, properties approximately identical to them)[19] are involved therein, and they will be expressed by general categorial terms – in the case of "water", e.g., by the term "liquid".

I already mentioned some assertions in Devitt & Sterelny (1999) and in other works by Devitt questioning the necessity of including descriptive components in a theory of reference borrowing or at least somewhat reluctantly conceding that the reference borrowing may comprise some *small* descriptive components, although Devitt avoids entering into this question. Thus, assuming that words express concepts, Devitt asserts that "the theory of reference borrowing places *very little* epistemic burden on the linguistically and conceptually competent […] There is, of course, room for argument about *just how little* an epistemic burden should be placed on the competent, but *we need not join this argument*" (Devitt 2006, 139; first and last emphases added).[20] And he hesitantly gives as an example of the descriptive component required for the reference borrowing of a word, or of its corresponding concept, that of the *type* of entity the referent is: "Perhaps there is some small epistemic burden on the person's conceptual competence so that the concept has some non-linguistic determiners; for example, perhaps the concept ‹Aristotle› has to be associated with the concept ‹human›" (Devitt 2006, 40). This example is basically the same as that given by Donnellan regarding the same name "Aristotle" and by Putnam with respect to the name "Quine".

According to the assertions by Putnam, partially by Donnellan, and more hesitantly by Devitt – despite Devitt's and Sterelny's asseverations in (1999) on the contrary – it is plausible that competent borrowers will have to associate some descriptive component which is true, or at least approximately true, of the referent of the terms, proper names and natural kind

---

[19] Some years after his (1975c), Putnam claimed that for two speakers to have acquired a natural kind term is not necessary that they associate with the term the same stereotype, but rather sufficient similar stereotypes (Putnam 1987, 271). It is to be assumed that this thesis would also be applicable to the semantic markers. However, according to the passage quoted above from Putnam (1975b, 203) I prefer to speak of "properties approximately identical" to the ones contained in the semantic markers instead of "properties sufficiently similar" to them.

[20] Concepts are the correlates in the language of thought – hypothesis accepted by Devitt – of the words of a natural language; thus words express concepts.

terms, which they borrow. In this regard the least questionable descriptive component is very general, i.e. the one concerning the *type* of entity referred to, which will be expressed by some *general categorial term* – and hence by the indefinite description formed with it. If the speaker is *completely* ignorant or wrong about the type of entity the referent is, it can be questioned that the borrower be a competent speaker.

Of course, the latter claim depends on what is required to be a competent speaker. As already said, in Devitt's and Sterelny's view in (1999) reference borrowing entails competence, but we can leave aside that specific view of competence and assume a more theory-neutral view of competence in a language, which they characterize as "the ability to produce and understand sentences with the sounds and meanings of that language" (Devitt & Sterelny 1999, 188; Devitt 2006, 201).

According to that theory-neutral view, our judgment on (lexical) competence depends on our conception of understanding and meaning. The authors who adopt a view of meaning strongly relying on a causal theory of reference will support a purely causal theory of competence. A view of that sort is proposed by Devitt and Sterelny, who identify the *sense* of a proper name mainly with "the property of designating its bearer by a certain type of causal link between name and bearer" (Devitt & Sterelny 1999, 67). Although hardly anyone else has shared that view of sense, this should be mitigated by Devitt's claim already quoted according to which "the theory of reference borrowing places very little epistemic burden on the linguistically and conceptually competent" (Devitt 2006, 139), but "very little" is still something. And although he does not want to deal with the question about what that "very little epistemic burden" should consist in, in the example he hesitantly gives, as indicated above, that "burden" concerning the proper name "Aristotle" is expressed by the general categorial term "human", which conveys the type of entity that Aristotle is. Another view of that sort, but different from the one held by Devitt and Sterelny, is the one embraced by advocates of the direct reference theory. However, even some of them also concede hesitantly that a competent speaker regarding the term "water" has to associate with this term the property of being a liquid (see Soames 2005, 184).

In fact, it is plausible that a competent speaker concerning the word "water" – i.e., who understands that word – associates with it at least the property expressed by the general categorial term "liquid", one regarding

the word "Aristotle", the property expressed by the general categorial term "human being", etc., or some approximately identical properties. If such a view is accepted one should also admit a certain sort of epistemic component in the notion of competence,[21] and that component is constituted by the property expressed by some general categorial term that conveys a general property of the referred entity, the type of entity it is. I should emphasize the character of centrality of those properties, since such comprehensive properties are more central than more specific ones. Thus, it is more central for the competence concerning the term "Quine" the property expressed by the term "human being" than the one expressed by the term "human being who was born on 25 June 1908 in Akron, Ohio". So the most central properties would be the most general or comprehensive properties,[22] which are the less susceptible to be questioned by the arguments of the ignorance-and-error type.

This view of competence, however, gives rise to some questions and, in particular, the following two. Firstly, since there are many general categorial terms that can be associated with a term – proper name or natural kind term – which express properties that are true or approximately true of its referent, the question arises regarding what to say about a speaker who associates with the term some of those properties, but not others. Let us assume that a speaker knows that Quine is a human being, but not a philosopher. From my point of view this speaker is competent insofar as he is knowledgeable about a general property that is true of Quine, although he is not as competent as other speakers that know that Quine is a human being and a philosopher. Competence, at least according to an epistemic view of

---

[21] Devitt has maintained that linguistic competence is "a piece of knowledge-how not knowledge-that" (see, e.g., Devitt 1981, 95-103; 1996, 52; 2006, 89-94 and 106, 2010, 142, n. 17 and 285; Devitt & Sterelny 1999, 173 ff.). Further to my foregoing considerations, the competence concerning proper names and natural kind terms must contain a modest knowledge-that, although in a different sense from the one meant by Devitt in his criticism of the knowledge-that's view of competence, according to which after "that" there should come "a sentence expressing something semantical about the language" (see Devitt 1981, 95).

[22] On this subject, someone could claim that the term "material object" is still more general than the examples of general categorial terms I have indicated above. But the information provided by the corresponding property is practically null, almost as null as that expressed by the so-called dummy sortals, such as "object", "thing", etc.

it, is a matter of degree. Thus, speakers who know many general properties that are true – or approximately true – of the referent of a term are more competent than other speakers who know a few or only one general property that is true – or approximately true – of the referent. Secondly, the question could be raised as to the necessity or sufficiency for the competence concerning a term of properties expressed by general categorial terms. On this matter my view is quite modest: the minimum necessary and sufficient condition for the competence about a term is expressed by some general categorial term that is true or approximately true of its referent.

At this point, it can be argued as follows. Speakers can be divided into different sorts. On the one hand, the grounders of a term, who associate descriptive components with the term to sort out the *qua*-problem concerning proper names and the two parts of the *qua*-problem regarding natural kind terms; on the other hand, the *competent borrowers* of the term, who associate with the term some general categorial terms that express very general properties that are true – or approximately true – of the referent, and that convey the type of entity referred to. Lastly, those speakers who, although having borrowed the term, are completely ignorant or wrong about the properties, even the most general ones, possessed by the referent of the term. Only the first two sorts of speakers are *competent*. Concerning the latter it could be claimed that even if they were to refer to an entity by the use of a term according to a purely causal theory of reference borrowing, they would have no idea whatsoever about the type of entity they refer to and so in this sense they have no idea as to what they refer to. Accordingly, they are not competent speakers concerning the term in question.[23]

Thus, my contribution to the debate concerning the theory of reference borrowing is that, adopting a moderately epistemic view of competence, at least the descriptive component required for the reference fixing of proper names in Devitt & Sterelny's theory in (1999) and the first descriptive

[23] A referee made the suggestion of not building the descriptive requirement into the theory of reference borrowing, but rather into the theory of what it is to be competent with a term. I could agree with this suggestion, but this is not the case in the theory of reference borrowing put forward by Devitt and Sterelny, the backbone of this paper, which joins both aspects. That is, according to those authors, the speaker who borrows the reference of a term – in a pure causal way –is a competent speaker concerning the term.

component needed to be associated with a natural kind term for its reference fixing in such a theory, i.e., some *general categorial term,* is also a requisite for the competence of reference borrowers. Therefore, as long as causal theorists consider borrowers as competent speakers, they should maintain a *descriptive-causal theory of reference borrowing*, which involves causal chains – or a causal network – in addition to some general categorial term, which is true or approximately true of the referent of the term.

Acknowledgments

This paper significantly modifies and further develops ideas put forward in Fernández Moreno (2016). The author is grateful to the comments made by two anonymous referees. This paper has been supported by the Spanish Ministry of Economy and Competitiveness in the framework of the research Project FFI2014-52244-P.

References

DEVITT, M. (1981): *Designation*. New York: Columbia University Press.

DEVITT, M. (1996): *Coming to Our Senses. A Naturalist Program for Semantic Localism*. New York: Cambridge University Press.

DEVITT, M. (2006): *Ignorance of Language*. Oxford: Clarendon Press.

DEVITT, M. (2008): Reference Borrowing: a Response to Dunja Jutronić. *Croatian Journal of Philosophy* 8, 361-366.

DEVITT, M. (2010): *Putting Metaphysics First. Essays on Metaphysics and Epistemology*. Oxford: Oxford University Press.

DEVITT, M. (2015): Should Proper Names Still Seem So Problematic? In: Bianchi. A. (ed.): *On Reference*. Oxford: Oxford University Press, 109-143.

DEVITT, M. & STERELNY, K. (1999): *Language and Reality*. *An Introduction to the Philosophy of Language.* Cambridge, Mass.: MIT Press, 2nd revised and enlarged edition; 1st edition, 1987.

DONNELLAN, K. (1972): Proper Names and Identifying Descriptions. In: Davidson, D. and Harman, G. (eds.): *Semantics of Natural Language*. Dordrecht: Reidel, 356-379.

FERNÁNDEZ MORENO, L. (2016): *The Reference of Natural Kind Terms*. Frankfurt am Main: Peter Lang.

HALE, B. & WRIGHT, C. (eds.) (1997): *A Companion to the Philosophy of Language*. Oxford: Blackwell.

JUTRONIĆ, D. (2008): Reference Borrowing and the Role of Descriptions. *Croatian Journal of Philosophy* 8, 349-360.

KRIPKE, S. (1980): *Naming and Necessity*. Oxford: Blackwell.

KRIPKE, S. (1986): A Problem in the Theory of Reference: the Linguistic Division of Labor and the Social Character of Naming. In: Cauchy, V. (ed.): *Philosophy and Culture. Proceedings of the XVIIth World Congress of Philosophy*. Montreal: Éditions de Beffroi, 241-247.

PUTNAM, H. (1975a): *Mind, Language and Reality*. (*Philosophical Papers*, Vol. 2). Cambridge: Cambridge University Press.

PUTNAM, H. (1975b): Explanation and Reference. In: Putnam (1975a), 196-214.

PUTNAM, H. (1975c): The Meaning of 'Meaning'. In: Putnam (1975a), 215-271.

PUTNAM, H. (1975d): Language and Reality. In: Putnam (1975a), 272-290.

PUTNAM, H. (1987): Meaning Holism and Epistemic Holism. In: Cramer, K. et al. (eds.): Theorie der Subjektivität. Frankfurt: Suhrkamp, 251-277.

PUTNAM, H. (1988): *Representation and Reality*. Cambridge: MIT Press.

PUTNAM, H. (2001): Reply to Michael Devitt. *Revue Internationale de Philosophie* 218, 495-502.

SOAMES, S. (2005): *Reference and Description. The Case against Two-Dimensionalism*. Princeton: Princeton University Press.

STERELNY, K. (1983): Natural Kind Terms. *Pacific Philosophical Quarterly* 64, 110-125.

THOMASSON, A. L. (2007): *Ordinary Objects*. Oxford: Oxford University Press.

Organon F 25 (4) 2018: 495-516

The Free Choice Principle as a Default Rule

DANIELA GLAVANIČOVÁ[1]

ABSTRACT: It is quite plausible to say that *you may read or write* implies that *you may read and you may write* (though possibly not both at once). This so-called *free choice principle* is well-known in deontic logic. Sadly, despite being so intuitive and seemingly innocent, this principle causes a lot of worries. The paper briefly but critically examines leading accounts of free choice permission present in the literature. Subsequently, the paper suggests to accept the free choice principle, but only as a default (or defeasible) rule, issuing to it a ticket-of-leave, granting it some freedom, until it commits an undesired inference.

KEYWORDS: Defeasibility – default rule – free choice permission – non-monotonic logic – paradox.

1. Introduction

The main topic of this paper is the free choice effect of a disjunctive permission. Let me start with some examples taken from the British National Corpus (BNC):[2]

[1] Received: 9 January 2018 / Accepted: 28 August 2018

✉ Daniela Glavaničová

Department of Logic and Methodology of Sciences
Comenius University in Bratislava
Faculty of Arts, Gondova 2, 814 99 Bratislava, Slovak Republic

e-mail: daniela.glavanicova@gmail.com

[2] The British National Corpus, version 3 (BNC XML Edition), 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

(1) You may sit down or stand just as you wish.
(2) You may exchange it or have your money refunded.
(3) You may copy a sound cassette or a video tape or disc for private use.
(4) You may also use a clean spoon or piece of paper.
(5) You may take mathematics with music or politics with personnel management.
(6) You may print/copy/delete either a subset or all of your oldest mail messages.

All these sentences allow an agent to freely choose between two or more options. This so-called *free choice permission* has been extensively discussed in the field of deontic logic. Hans Kamp (1973, 57) used the following example to introduce the paradox of free choice permission:

> (BC) *You may go to the beach or go to the cinema*, I almost told my son Michael. But thought better of it, and said: (B) *You may go to the beach.* Boys shouldn't spend their afternoons in the stuffy dark of a cinema, especially not with such lovely weather as to-day's.

Intuitively, the latter permission is entailed by the former, but not *vice versa*. However, Standard Deontic Logic (SDL; normal propositional modal logic with serial accessibility relation) tells the opposite story. Let (BC) be represented as $P(b \vee c)$ and (B) as Pb (where P is a deontic operator of permission). In SDL Pb implies $P(b \vee c)$ (since the operator of permission is closed under classical consequence), but $P(b \vee c)$ does not imply Pb. However, incorporating the intuition that (BC) implies (B) by adding the corresponding principle into SDL results in a logical apocalypse.

The principle in question is well-known as the free choice principle:

(FCP) $\quad P(\varphi \vee \psi) \rightarrow P\varphi \wedge P\psi$

There has been a lot of pessimism surrounding the intuitively plausible and practically useful[3] FCP. To some extent, the pessimism is justified. Sven

[3] Consider, for instance, its usefulness related to agency. While it may not be clear how to obey some disjunctive commands or how to exercise disjunctive permissions or

Ove Hansson (2013) nicely sums up implausible formulas that can be subsequently derived. One of them has been derived using the FCP and the substitution of equivalents only.[4] For Hansson, this indicates that FCP may be faulty in itself. A different approach suggests the problem should be solved within the domain of pragmatics. Yet different stance was taken by Zimmermann (2000) or Anglberger, Faroldi & Korbmacher (2016), who abandoned the substitution of equivalents, allowing only for the substitution of hyperintensional equivalents. Note, however, that Zimmermann restricted the validity of free choice principle to cases where the person granting permission has the needed authority.

The main idea of the present paper is simple: Let us add the troublesome-yet-intuitive FCP, but only as a *default rule*. To this purpose, a nonmonotonic framework of adaptive logics will be used. Of course, there are many others options. What motivates the choice of adaptive logic is the dynamic character of its logic (see Beirlaen, Straßer & Meheus 2013, 296-298 and Goble 2013a, 338-339). It seems that free choice effect can be cancelled in the process of communication, and again "resurrected" afterwards. As we will see later on, this nicely corresponds to the idea of "marking" in adaptive logics. Furthermore, FCP will be accepted as a *rule*, not as an *axiom*. While in general, if one has a logic which already has Modus Ponens (MP), it makes little difference whether we opt for an axiom in the form of implication and MP, or for a specific rule. Yet some motivations can be provided. One trivial reason is that it is a natural option as soon as one uses adaptive logics. An independent reason: it has some advantages concerning one of the implausible consequences of adding FCP into SDL, Hansson's implausible result 4, and similar inferences (again, this will be explained later on, when we'll have all building blocks needed for the explanation at our disposal). Another important feature of the account to be proposed is that "strong" free choice permission will be distinguished from standard SDL permission.[5] Following the notation suggested by Hansson, free choice permission between A and B will be written as $Pc(A \vee B)$ while

rights, FCP suggests a solution at least for permissions. Hansson (2013, 209) discusses related phenomena in the domain of commands.

[4] By the substitution of equivalents the substitution of classical logical equivalents will be meant throughout the paper.

[5] Alternatively, one could use a non-truth-functional disjunction.

SDL permission will have its standard notation. Only disjunctive non-modal formulas can occur within the scope of free choice permission, and it will be impossible to derive a free choice permission from formulas that have no occurrence of free choice permissions (in other words, free choice permissions will occur as premises rather than consequences).

The structure of the paper is as follows. First, I will recapitulate reasons why SDL should not be enriched with FCP as it stands (Section 2). Second, I will consider some approaches known in the literature and formulate some objections against them (Section 3). I will argue that FCP should be treated as a defeasible principle: it can be fruitfully employed, but its use can be also reasonably cancelled (Section 4). Subsequently, I will explain the connection between defeasibility and non-monotonic logics very briefly (Section 5). Finally, I will introduce and defend FCP as a default rule, showing how implausible results can be avoided and defeasibility maintained employing adaptive logics (Section 6) and conclude the paper (Section 7).

2. FCP meets SDL: implausible consequences

As Hansson (2013, 207) puts it, although the free choice postulate "seems innocuous when presented in connection with a permitted choice, in combination with other deontic postulates it gives rise to a whole series of implausible results". Hansson discusses some of them (see Hansson 2013, 207-208):

(IR1) $OA \rightarrow O(A \wedge B)$
(IR2) $OA \rightarrow PB$
(IR3) $PA \rightarrow PB$
(IR4) $PA \rightarrow P(A \wedge B)$

The derivation of IR1 requires the substitution of equivalents and interdefinability ($OA \leftrightarrow \neg P \neg A$). IR2 requires the substitution of equivalents, $OA \rightarrow PA$ and $OA \rightarrow O(A \vee B)$. IR3 requires $O(A \wedge B) \rightarrow OA$ and interdefinability. IR4 has been derived using only FCP and the substitution of equivalents. Another problematic inference concerning FCP is the Hansson's (2013, 218) *Vegetarian's Free Lunch*, which goes as follows: "You may

have a meal with meat or a meal without meat. Therefore you may either have a meal and pay for it or have a meal and not pay for it."

Should we abandon the principle, as numerous implausible consequences suggest? Should we wave goodbye to SDL and accept the principle as an infallible logical rule, as the intuitive plausibility of the principle suggests? Is there a middle way between these two extremes? Or should we abandon semantics and divert to pragmatics?

3. FCP: state of art

Let us first have a look at the approaches present in the literature. Hansson (2013, 209-218) lists five main types of proposed solutions to the problem of free choice permission, which can be divided into two categories, semantic and pragmatic. Hansson claims that the second approach is pragmatic and the rest belongs to the semantic category.

The first approach, the *mistranslation of or*, claims that the problem of free choice permission arises due to a mistranslation of *or* from some natural language into some logical language. When we say "You may A or B", this *or* is not a truth-functional disjunction, but a connective for contracted sentence parts. The above sentence is a contraction for *You may A and you may B* (so-called dummy connective approach). We should thus represent it as $PA \wedge PB$, rather than as $P(A \vee B)$. However, this leaves at least one question open: which part of $PA \wedge PB$ corresponds to "or"? It is not transparent how we acquired this formalization. Alternatively, we should represent it as $(\forall x)(x=A \vee x=B \rightarrow Px)$, what is equivalent to $PA \wedge PB$ (so-called checklist conditional approach). This is the approach advocated by Makinson (1984).

The second main approach goes by the name *conversational implicature* and it suggests that free choice effect is not inherent in the language, but implied by the context of utterance, thereby being a pragmatic, rather than semantic phenomenon. The Gricean mechanism is thus invoked to explain the free choice effect. The predominant pragmatic view is to understand free choice inferences as a sub-species of scalar implicatures (see Kratzer & Shimoyama 2002; Chemla 2009; Singh et al. 2016, among others). Interestingly, Kratzer & Shimoyama (2002) consider the free choice inference as an implicature *because* of its cancellability.

The third approach can be entitled *a hidden operator approach*. This approach understands *or* as ambiguous between truth-functional disjunction and a connective including *and you may choose which*. Free choice effect is thus inherent in one of the meanings of *or*. Alternatively, free choice effect can be inherent in the syntactic structure (surely not in the surface structure, but possibly in the logical form of the sentence). Such an account has been called syntax-based. Hans Kamp (1973) takes this route.

The fourth approach on Hansson's list is called *free choice operators*. Hansson explains that according to this approach, "[t]he 'or' of free choice permission follows other logical laws than those of ordinary permission" (Hansson 2013, 210). This claim is a bit misleading, since what is proposed here is not specific disjunction, but specific permission.

The fifth, Hansson's own approach, *the impossibility of single-sentence representation*, claims that free choice permission is a property of the sets of action describing sentences, rather than a property of disjunction of these sentences. In "You may A or B", free choice permission is understood as a property of the set {A, B}. Yet, as Hansson (2001, 131) notes, the Makinson's (1984) checklist conditional approach satisfies this criterion too. For this reason, it is not clear why is it listed as a different solution.

A hyperintensional approach was suggested by Zimmermann (2000) and also by Anglberger, Faroldi and Korbmacher (2016). Unfortunately, hyperintensional approach is not mentioned in Hansson's list of proposed solutions. The approach is closely related to mistranslation of 'or', yet it cannot be subsumed under this category as specified by Hansson. It is also related to the fifth approach. Anglberger, Faroldi and Korbmacher propose the exact truth-maker semantics, which makes the free choice principle valid. Zimmermann proposes that a disjunctive permission should be analysed in terms of a special, non-Boolean disjunction (disjunction as a conjunctive list of epistemic possibilities). His approach has another distinctive feature: while the previous proposals were trying hard to validate the free choice postulate (what was, after all, the original goal), Zimmermann denies that the free choice postulate is valid. Rather, he claims that free choice effect *does not always come about*, though it sometimes does. In particular, free choice effect arises when the speaker is an authority on issue in question. For any context c and property P the speaker is an authority on P in c iff the speaker knows P's extension in c. Zimmermann gives us some examples of such authorities, be it a legal advisor or someone who

has just read the book of rules. Extreme examples are performative uses (saying so makes it so), e.g. a father giving a permission to his child. Zimmermann presents us also with his peculiar solution to the challenge of defeasibility:

> Obviously, cancellations of the choice effect are no problem for the present approach. Indeed, by uttering a sentence like [Mr. X may take a bus or taxi, but I don't know which] the speaker explicitly reveals that she is not an authority – if not remembering is taken as indication of a lack of knowledge. (Zimmermann 2000, 287)

On the contrary, the approach suggested by Anglberger, Faroldi and Korbmacher has no place for cancellations or situations where the free choice effect does not come about. This is so because the free choice permission is incorporated straight into the proposed semantics. Hyperintensional approach completes the list of main proposals, yet for sure, not all of existing proposals could be outlined here.

Let me now assess the presented approaches and point to some of their drawbacks. The main problem I see with the first approach is closely related to the very motivation for the solution which will be suggested in this paper. The issue is that if "You may A or B" is unambiguously translated into a logical formulation equivalent to "You may A and you may B", no weakening and no defeating of the free choice effect is possible. Yet as the next section argues, there are such cases of weakening or defeating of the free choice effect. Also, as shortly indicated above, the first version of this approach leaves no clues *why* we tend to use *or* instead of *and*, if *and* is what we originally meant. Neither is it clear what part of the formulation corresponds to the original *or*. In contrast, the checklist conditional approach suggests an elegant and transparent formulation: not only is disjunction preserved, it is also clear how it leads us to a conjunctive meaning. A disadvantage is that we have to leave the propositional language, which is so commonly used in deontic logic. This may be seen as a too high price to pay, given that the only gain is the apparent transparency of the free choice disjunction.

The second approach locates free choice permission in the realm of pragmatics. Importantly, it allows us to *derive* the free choice effect, and to subsequently *defeat* the very same effect. One trouble with the so-called

scalar implicature view is that the account of free choice inference requires distribution over disjunction, while there are free choice inferences related to abilities, but distribution over disjunction $\Diamond(A \vee B) \rightarrow \Diamond A \vee \Diamond B$ is not generally valid for them (see Nouwen 2018). Moreover some recent findings suggest that there are considerable differences between scalar implicatures and free choice inferences (see Chemla & Bott 2014, Tieu et al. 2016). As Willer (2017) rightly notes, this is, however, not a sufficient evidence for showing that the phenomenon is not pragmatic. Now a question whether the phenomenon should be addressed by semantics or by pragmatics is a serious question for philosophy.[6] Yet the present question (for the logician) is rather how to capture these derivations and cancellations *in logic*. This is something the pragmatic solution leaves open (or worse, it leaves us with the literal meaning, the original unhelpful and implausible SDL formalization of "You may A or B" as $P(A \vee B)$). My reply to the semantic/pragmatic localization of the phenomenon is as follows: while there are tests for finding out whether some phenomenon is semantic or pragmatic, they seem to be inconclusive. Moreover, even some empirical data go against the implicature view: if the free choice inference was a scalar implicature, the restricted time would result in a decreased rate of free choice responses. However, this hypothesis was falsified (Chemla & Bott 2014). This result corresponds to the linguistic intuition that we in fact don't execute complicated inferences to derive the free choice effect. This evidence gives us some reasons to deny that free choice is a pragmatic phenomenon. But more importantly, and in line with Willer (2017), *whether the phenomenon is semantic or pragmatic, the logician may suggest a logic for this phenomenon*. If the phenomenon is semantic, s/he might claim that the suggested logic captures *the literal meaning* of the free choice permission. If the phenomenon is pragmatic, s/he might claim that the suggested logic captures *the communicated meaning*, or the implied content, or utterances containing free choice permission. Importantly, implicatures in general can be rather smoothly analysed as default rules, for their cancellability is an acknowledged phenomenon. In other words, even if the phenomenon was pragmatic, the solution I am about to offer is a natural choice.

Hidden operator approach postulates lexical or syntactic ambiguity. Disjunctive permission is thus once analysed in the free-choice manner,

---

6 Thanks to one of the *Organon F* reviewers for pressing me to address this issue.

once in the classical manner (i.e. without the free choice effect). One problem is, however, how to delineate cases with free choice reading from cases without it. A similar worry as the one related to the first approach applies here: If we disambiguate a sentence as having the free choice reading, the free choice effect cannot be subsequently cancelled.

As regards the fourth approach, various operators for the free choice permission lead to various implausible results (see Hansson 2013, 214-217). All of such solutions share a common assumption, which Hansson believes to be the root of inadequacy of such approaches, namely, the single sentence assumption: "Free choice between *a* and *b* can be represented as a property of a single sentence, namely $a \vee b$" (Hansson 2013, 218). Hansson claims that this assumption leads to a troubling consequence: If $a \vee b$ and $c \vee d$ are equivalent, then there is a free choice permission between *a* and *b* iff there is a free choice permission between *c* and *d*. This leads to absurd consequences, such as the Vegetarian Free Lunch example (recall: you have a meal with meat or without meat; therefore you may either have a meal and pay for it or have a meal and not pay for it).

Hansson is surely right that this is absurd. Yet, is he really right that the trouble is a consequence of the single sentence assumption? There is an alternative that free choice permission in fact creates (hyper)intensional context, and does not allow for unrestricted use of the extensionality principle.

Hansson does not seem to take this possibility into account and suggests that free choice permission should be represented as a property of the set of action-describing sentences, because

> (free choice) permission to perform either *a* or *b* is not a function of a single sentence $a \vee b$ but a function of the two sentences *a* and *b*. It is a function of two variables, not one. Similarly, (free choice) permission to perform *a*, *b*, or *c* is a function of three variables, etc. (Hansson 2013, 218)

Unfortunately, he does not present us with much details of his account, just with the main idea.

Zimmermann claims that cancellations of the choice effect are no problem for his approach. His explanation of the occasionally fading free choice effect seems to have some rationale. Yet this explanation is not satisfactory

enough: Zimmermann admits that free choice effect is sometimes cancelled, but this would be possible only if the free choice effect *would have been originally present*: by some utterance, we are cancelling something what was previously uttered or implied by our utterance. However, in Zimmermann's account the very notion of cancellation is not applicable. Cancellations are thus a problem for this approach. The proposal of Anglberger, Faroldi and Korbmacher validates the free choice principle and thereby inherits the main disadvantage of the first approach: free choice effect is present whenever a disjunctive permission is, and cannot be cancelled.

4. FCP and defeasibility

As indicated, if the pragmatic approach is right, free choice effect should be defeasible. Indeed, this seems to be acknowledged as an obvious fact in this vein of literature (see for instance Tieu et al. 2016, Kratzer & Shimoyama 2002). Semantic approaches seem to count with this phenomenon too (see Zimmermann 2000, Anglberger, Gratzl & Roy 2015).

Let us consider some examples where the free choice effect is weakened or defeated. Hans Kamp's thoughts contain one such example, when he wanted to utter (BC), leading to the free choice effect, but uttered (B) instead (see Section 1). Alternatively, one can adjust the original example in the following way (three dots stand for Kamp's contemplative moment):

(7) *You may go to the beach or go to the cinema. ...But first ask your mother.*

(8) *You may go to the beach or go to the cinema. ...But you don't have enough money for a cinema ticket and I won't give you any.*

(9) *You may go to the beach or go to the cinema. ...But I don't know which one.*

(10) *You may go to the beach or go to the cinema. ...But there are sharks in the ocean, don't go to the beach.*

(11) *You may go to the beach or go to the cinema. ...But the cinema is under the reconstruction. It is closed this month.*

(12) *You may go to the beach or go to the cinema. ...But not only to the closest cinema to our house.*

(13) *You may go to the beach or go to the cinema. ...But don't leave the town!*

Different kinds of defeasibility occur in (7)–(13). In (7), the father in question is not the sole normative authority (this phenomenon was discussed also by Zimmermann 2000) and though his permission is granted, permission from someone else (e.g., the child's mother) is needed before exercising disjunctive permission from (7). In (8), disjunctive permission is granted, but the practical possibility of realisation of one "option" is questioned. How could (9) possibly happen? One such scenario would be the following: The father is no normative authority in this respect, but he still remembers some permission has been given by someone else (e.g., the child's mother again). This example is different from others at least in two respects: first, epistemic modality is intertwined with deontic modality; second, while the first sentence has the same form as in other examples (i.e. it is a disjunctive permission), it is indicated that the free choice effect was not present at all: the father doesn't know *which* one. This suggests that *only one* of those two actions has been permitted (admittedly, by someone else), but the father doesn't remember which one. (How would theories which suggest that "or" is just mistranslated into the logical language as disjunction, whilst the real meaning is conjunctive, reply to this sort of examples?)[7] In (10), free choice is defeated, though it surely was present at the time of uttering of the sentence expressing disjunctive permission. In (11), both options are granted, but as in (8) the practical possibility of realisation of one option has been challenged, here the practical impossibility of such realisation is suggested. In (12) and (13), restrictions are imposed upon the admissible ways of realizing the permission in question. These may lead to cancellations too: e.g., imagine a situation where there's no cinema in the town, or where the nearest cinema is closed or too expensive.

[7] As one of the reviewers pointed out, it can be claimed that (9) is not even a free choice permission situation. Yet if we imagined it uttered in some dialogue, after uttering, the addressee would understand it as free choice permission. It is only after the latter sentence that this prescriptive *and* free choice reading would be shown implausible, and thus, in a sense, defeated.

Interestingly, free choice inferences can be even more dynamic: free choice effect can appear, disappear, reappear… Consider the following dialogue:

Father: *You may go to the beach or to the cinema.*
Father: *But don't leave the town!*
Son: *But the cinema in our town is closed.*
Father: *Go to the beach then.*
Son: *I was there yesterday.*
Father: *Ok, you may leave the town.*
Father: *But first, ask your mother!*

The free choice effect has appeared in the first replica, disappeared in the subsequent communication, and reappeared by the allowance to leave the town, started to fade out again by the father not being the sole authority in the present case.

For sure, Hansson's phenomenon of defeasibility is not restricted to disjunctive permissions. Of course, almost anything can be defeated in the flow of communication, not only the free choice effect. But there is one important feature pertaining to disjunctive permissions: one can defeat or weaken just one conjunct of the consequent of FCP after stating only its antecedent. This means that the free choice effect itself is defeated or weakened. Also, defeasibility may occur in the connection to various rules and phenomena within deontic logic and normative reasoning (cf. for instance the paper Mullins 2016 claiming that rights should be treated in terms of default logic, or motivations for introducing any non-monotonic deontic logic, since, as it will be explained shortly, the two are closely related). However, what concerns me is solely the free choice principle, and how we should treat it. My suggestion is that it is indeed a very useful and plausible principle, but some care should be taken, and it should be accepted only as a default rule, not as an infallible principle. What is also clear is that SDL and similar normal modal logics are static and monotonic: what was once permitted stays permitted. In other words, the addition of new information cannot defeat the previous consequences of a (normative) system. We thus need some deontic logic that is not static and monotonic.

5. Defeasibility and non-monotonic logics

The problem of defeasibility is being almost uniformly solved with the help of non-monotonic logics. Let me explain this a bit. On the one hand, classical logic is monotonic: suppose we have some premise set $\Pi_1 = \{p_1,...,p_n\}$ and $c \in Cn(\Pi_1)$ (i.e. c is a logical consequence of Π_1). Now imagine that we gain further information p_{n+1}, so we add it to our "knowledge base", thus creating the premise set $\Pi_2 = \{p_1,...,p_n, p_{n+1}\}$. Obviously, $\Pi_1 \subseteq \Pi_2$ holds and so $Cn(\Pi_1) \subseteq Cn(\Pi_2)$ holds classically too. It is thus impossible that $c \notin Cn(\Pi_2)$.

Monotonic logic is entirely satisfactory if we want to derive consequences of complete, static and consistent information about some domain. However, this is usually not the case in this world of imperfectness. The field of artificial intelligence aims to deal with reasoning from incomplete or inconsistent knowledge bases, and because of this, non-monotonic reasoning is widely studied in the field. Formally, operator of logical consequence *Cn* is non-monotonic, if for some sets Π_1, Π_2 such that $\Pi_1 \subseteq \Pi_2$ it holds that $Cn(\Pi_1) \not\subseteq Cn(\Pi_2)$. So our set $Cn(\Pi_2)$ can possibly miss *c*.

Non-monotonicity is a fundamental feature of default reasoning. The most influential paper in the field is surely Reiter's 1980 paper *A Logic for Default Reasoning*. As Reiter explains, despite the fact that we do not have total knowledge about some domain, we must sometimes draw conclusions based on our incomplete information. Default reasoning arises on this ground and it amounts to an inference of the following form: in the absence of any information to the contrary, assume...

Deontic extensions of logics for default reasoning have been introduced mainly because of the obvious existence of so-called normative conflicts in natural language (a normative conflict obtain when $O\varphi \wedge O\neg\varphi$ holds for some φ). While normative conflicts are quite common in natural language, they make standard deontic logicians feel uneasy. First of all, an occurrence of normative conflict in SDL leads to inconsistency. Furthermore, it leads to so-called deontic explosion. Another troublesome consequence is Chisholm's famous paradox. All these worries motivated deontic logicians to devise non-monotonic deontic logics (see Lou Goble's chapter on normative conflicts as evidence – Goble 2013a). A non-monotonic approach allows us to retain most of the standard principles and still avoid the most troublesome consequences.

As suggested above, I believe that situation with the free choice problem is similar in this respect to the situation with normative conflicts. On the one hand, we have plausible principles of standard deontic logics. On the other hand, we have the troublesome (but still plausible) free-choice principle, which is incompatible with these principles and which can be defeated.

6. FCP meets non-monotonic logic

Following Reiter's pioneering 1980 work, one can reformulate FCP in natural language as follows:

(FCP*) If it is permitted that φ or ψ, then it is *usually* permitted that φ and permitted that ψ.

It has to be specified in advance what is understood by a disastrous consequence (contradiction is the prime example of a disastrous consequence, normative conflicts can be listed as another example – what else?). Consequently, any use of FCP* that leads to a disaster will be cancelled. FCP* will be at hand anytime, helping us to generate consequences, but non-monotonicity will help us to avoid logical disasters.

Now various non-monotonic deontic logics can be used. For the present purposes it matters little whether one opts for Horty's default logic (Horty 1993; 1997), or for adaptive logics, or for some other framework. I will use adaptive logics as the framework for treating free choice permissions. As we will see very soon, this choice can be motivated by similarities between dynamic character of the proof theory of adaptive logics and dynamic character of free choice inferences.

Adaptive logic is an interesting framework for default reasoning, developed mainly by Diderik Batens (see Batens 2007, Batens & Haesaert 2002, Goble 2013a and 2013b). In general, an adaptive logic *AL* is a triple ⟨*LLL*, *Ω*, *Strategy*⟩. *LLL* is so-called lower limit logic, which is reflexive, transitive, monotonic, compact, has characteristic semantics and contains classical logic. *Ω* is a set of abnormalities, which is *LLL*-contingent (neither abnormalities nor their formal negations are theorems of *LLL*) and contains at least one logical symbol. *Strategy* is a method how to evaluate proofs

where 'abnormal' consequences have been derived. Most widely used strategies are *reliability strategy* and *minimal abnormality strategy*. The proof theory of adaptive logics consists of three generic rules, namely

> a simple rule of premise introduction, PREM, and a rule RU that accepts unconditionally all inferences valid in LLL. And then the conditional or provisional rule RC that is characteristic of adaptive logics. (Goble 2013b, 9)

The key idea behind the proof theory of adaptive logic is *marking*. Some lines of proofs are marked, some are not. If a line is marked, formula occurring on it is no longer derivable. Yet, the very notion of derivability is unstable as "marks may come and go" (Batens 2007, 8).

Adaptive logics have their deontic versions (see Beirlaen, Meheus & Straßer 2013, Goble 2013a, 2013b, Van De Putte, Beirlaen & Meheus 2018). As already stated, an adaptive logic *AL* is a triple $\langle LLL, \Omega, Strategy \rangle$, where *LLL* is so-called lower limit logic, Ω is a set of abnormalities, and *Strategy* deals with problems, for instance, with inconsistencies. Nothing precludes the use of some deontic logic as *LLL*, if it is a reflexive, transitive, monotonic and compact logic, which has characteristic semantics and contains classical logic. For instance, we can use deontic extensions of classical propositional logic, such as SDL. Since adaptive deontic logics are used mostly to account for normative conflicts, their crucial aim is to avoid any form of deontic explosion and to account for some intuitive arguments that are usually problematic for deontic logics for normative conflicts. Though this is not the aim of the present paper, some inspiration can be drawn from the way in which are these systems introduced:

> In general, adaptive logics are a type of dynamic, non-monotonic system of reasoning designed to apply problematic rules, such as aggregation or distribution, provisionally. A use of the rule is accepted until it makes trouble, as gauged against a specified class of abnormalities, at which point, but only at that point in context, it is rejected. (Goble 2013a, 338)

What potentially problematic rules come into play in our case? Surely, the free choice postulate is such a rule.

The adaptive logic employed will be entitled $SDL_c^m = \langle SDL_c, \Omega_c, m\rangle$ (following the notation from Van De Putte, Beirlaen & Meheus 2018). SDL_c is *SDL* with a dummy operator *Pc* for free choice permission. As the *prima facie* obligation in the work quoted, free choice permissions cannot be derived from other free choice permissions. One constraint is that we allow only disjunctive formulas to be in the scope of *Pc*. Ω_c is specified by the logical forms $Pc(A \vee B) \wedge \neg PA$ and $Pc(A \vee B) \wedge \neg PB$; m stands for the strategy employed: minimal abnormality. It needs to be said that minimal abnormality is not the simplest strategy available and its precise definition is rather complex. However, as we will see later on, it has some advantages over the simpler reliability strategy. Informally, "we have sufficient reasons to infer A [if] every minimally abnormal way of interpreting the current proof stage will make A true" (see Van De Putte, Beirlaen & Meheus 2018, section 3).

To capture defeasibility of the free choice effect of disjunctive permission in terms of adaptive logic, rules of the type RC are at our disposal. In our case: From the free choice permission $Pc(A \vee B)$ infer PA (PB), under the constraint that none of the abnormalities in $\{Pc(A \vee B) \wedge \neg PA, Pc(A \vee B) \wedge \neg PB\}$ be derivable from Γ:

> If $Pc(A \vee B) \vdash_{LLL} PA \vee (Pc(A \vee B) \wedge \neg PA)$, then $Pc(A \vee B)$ implies PA unless $Pc(A \vee B) \wedge \neg PA$ is derivable.
> If $Pc(A \vee B) \vdash_{LLL} PB \vee (Pc(A \vee B) \wedge \neg PB)$, then $Pc(A \vee B)$ implies PB unless $Pc(A \vee B) \wedge \neg PB$ is derivable.

To demonstrate that these rules work as expected, let me start with constructing a proof of PA and PB from the premise $Pc(A \vee B)$ (to establish the validity of the formal representation of the free choice postulate in an adaptive logic):

1. $Pc(A \vee B)$	-	PREM	Ø
2. $PA \vee (Pc(A \vee B) \wedge \neg PA)$	1	RU	Ø
3. PA	1,2	RC	$\{Pc(A \vee B) \wedge \neg PA\}$
4. $PB \vee (Pc(A \vee B) \wedge \neg PB)$	1	RU	Ø
5. PB	1,4	RC	$\{Pc(A \vee B) \wedge \neg PB\}$

To show that defeasibility really works here, it is needed to add information $\neg PA$ as a premise (or to derive it):

6. $\neg PA$	-	PREM	Ø
7. $Pc(A \vee B) \wedge \neg PA$	1,6	RU	Ø

Since the formula of line 6 is so-called minimal Dab-formula that is derived on an empty condition, any line with this formula in conditions is to be marked. Because of this, the line 2 is to be marked (✓ is standardly written in front of the marked lines).

Can we go on with and defeat also PB? Having a different notation for free choice permission and for standard permission, we can do it without deriving contradiction. Yet if we wished to weaken this ability of defeating (i.e., quite plausibly claiming that $Pc(A \vee B)$ is inconsistent with having both $\neg PA$ and $\neg PB$), we can add an unconditional rule $Pc(A \vee B) \rightarrow PA \vee PB$. Be it as it may, the free choice effect can be easily cancelled within this framework, without thereby having a contradiction in the system.

Let me now motivate the employed strategy shortly. Inferring the free choice effect is consistent with adding "but not both". For instance, we might be told in a hotel restaurant that "You may have a cake or an ice cream as a dessert". Now sadly for a greedy person, "but not both" reading is usually assumed. An alternative reliability strategy would not allow us to have $PA \vee PB$ derived from the free choice permission $Pc(A \vee B)$ if $\neg PA \vee \neg PB$ (i.e., "but not both") is assumed. Consider the following proof:

1. $Pc(A \vee B)$	-	PREM	Ø
2. $PA \vee (Pc(A \vee B) \wedge \neg PA)$	1	RU	Ø
3. PA	1,2	RC	$\{Pc(A \vee B) \wedge \neg PA\}$
4. $PB \vee (Pc(A \vee B) \wedge \neg PB)$	1	RU	Ø
5. PB	1,4	RC	$\{Pc(A \vee B) \wedge \neg PB\}$
6. $PA \vee PB$	3	RU	$\{Pc(A \vee B) \wedge \neg PA\}$
7. $PA \vee PB$	5	RU	$\{Pc(A \vee B) \wedge \neg PB\}$
8. $\neg PA \vee \neg PB$	-	PREM	Ø
9. $(Pc(A \vee B) \wedge \neg PA) \vee (Pc(A \vee B) \wedge \neg PB)$	1,8	RU	Ø

Now reliability strategy has it that (Van De Putte, Beirlaen & Meheus 2018, Section 3) "a line is marked whenever its condition contains an abnormality that is a disjunct of a minimal Dab-formula that has been derived in the same proof." A minimal Dab-formula is contained in the line 9. This means

that according to the reliability strategy, lines 3, 5, 6 and 7 are marked as unreliable. However, this is not plausible: we want to keep the possibility to have at least some dessert! In other words, $PA \vee PB$ should be derived. Minimal abnormality allows us to have this result. Every minimally abnormal way of interpreting the current proof stage suggests that just one of the two abnormalities in question holds (either $Pc(A \vee B) \wedge \neg PA$ or $Pc(A \vee B) \wedge \neg PB$). But whichever of them holds, we can derive $PA \vee PB$, and because of this, this formula is derived (and we will have our dessert).

Finally, let us have a look on implausible results mentioned in Section 2 and see whether their derivation can be blocked within the present proposal. Note that what will block implausible results is not the non-monotonic logic, but the very fact that there are two kinds of permission: free choice permission of the form $Pc(A \vee B)$ which is given rather as an input than as an output, and Pφ of SDL. Because of this, I will leave derivations (with little amendments) as they were shown in Hansson (2013), and explain which of their steps will fail under the present proposal (strictly speaking, any line with free choice principle will fail, as it is not an axiom in the adaptive logic).

Derivation of (IR1) $OA \rightarrow O(A \wedge B)$,

1. $P(\neg A \vee \neg B) \rightarrow P\neg A$
2. $P\neg(A \wedge B) \rightarrow P\neg A$
3. $\neg P\neg A \rightarrow \neg P\neg(A \wedge B)$
4. $OA \rightarrow O(A \wedge B)$

is based on the equivalence between $P(\neg A \vee \neg B)$ and $P\neg(A \wedge B)$. This equivalence still holds under the present proposal, but $P(\neg A \vee \neg B)$ is clearly not a free choice permission, so the first line cannot be derived. On the other hand, if there were $Pc(\neg A \vee \neg B) \rightarrow P\neg$A in the first line, the equivalence with $P\neg(A \wedge B)$ cannot be assumed.

Derivation of (IR2) $OA \rightarrow PB$,

1. $P(A \vee B) \rightarrow PB$
2. $O(A \vee B) \rightarrow P(A \vee B)$
3. $O(A \vee B) \rightarrow PB$
4. $OA \rightarrow O(A \vee B)$
5. $OA \rightarrow PB$

is based on non-free choice disjunctive obligation seen as implying the free choice disjunctive permission. Again, while $O(A \vee B) \rightarrow P(A \vee B)$ holds, $O(A \vee B) \rightarrow Pc(A \vee B)$ does not. If we added "free choice obligations" into the language, the principle $Oc(A \vee B) \rightarrow Pc(A \vee B)$ would be correct. However, $OA \rightarrow Oc(A \vee B)$ would fail, as we cannot introduce the choice between obligations of A and B from the obligation of A.

Derivation of (IR3) $PA \rightarrow PB$,

1. $O(\neg A \wedge \neg B) \rightarrow O\neg A$
2. $O\neg(A \vee B) \rightarrow O\neg A$
3. $\neg O\neg A \rightarrow \neg O\neg(A \vee B)$
4. $PA \rightarrow P(A \vee B)$
5. $P(A \vee B) \rightarrow PB$
6. $PA \rightarrow PB$

again rests upon the conflation of two kinds of permission. The line 4 cannot be derived with free choice permission, which is, however, needed to derive (something similar to) the line 5.

Derivation of (IR4) $\rightarrow P(A \wedge B)$,

1. $P((A \wedge B) \vee (A \wedge \neg B)) \rightarrow P(A \wedge B) \wedge P(A \wedge \neg B)$
2. $PA \rightarrow P(A \wedge B) \wedge P(A \wedge \neg B)$
3. $PA \rightarrow P(A \wedge B)$

rests upon the extensionality of free choice permission, which, however, fails for this kind of permission. Another important thing is related to the free choice principle figuring as a rule rather that as an axiom. Even if one opted for adaptive logics with the free choice permission obeying the extensionality principle, thereby being able to substitute PA for $P((A \wedge B) \vee (A \wedge \neg B))$ in $P((A \wedge B) \vee (A \wedge \neg B)) \rightarrow P(A \wedge B) \wedge P(A \wedge \neg B)$, one would not be able to list the free choice principle in the first line: $P((A \wedge B) \vee (A \wedge \neg B))$ would have to be listed as a premise first, (provisionally) granting the permission of both A with B and A with $\neg B$. Under this supposition, the conclusion would be much less controversial (cf. also the open reading of permissions in Anglberger, Gratzl & Roy 2015).

7. Conclusion

The main topic of the present paper is disjunctive permission and its free choice effect. As is well-known, the addition of so-called free choice principle into SDL results in many troubles. Yet the principle itself seems to be very plausible and useful. Because of this "dilemma", the paper opted for a middle way: to accept the principle, but only as a default rule. This suggestion was further motivated by several examples of how the free choice effect can be easily defeated in the subsequent communication, but also by discussing and evaluating accounts formulated in the literature. After that, the paper explained that the phenomenon of defeasibility in natural language is standardly being solved in terms of non-monotonic logic. Finally, the paper defined an adaptive deontic logic SDL_c^m and showed how free choice effect can be derived and cancelled within this logic, and how implausible consequences can be avoided.

Acknowledgments

This work was supported by the Slovak Research and Development Agency under the contract no. APVV-17-0057 and by the VEGA grant no. 1/0036/17. The paper is based on my talk at XX. Slovak-Czech Symposium on Analytic Philosophy and my master's thesis. First and foremost, I would like to thank Marián Zouhar for being a perfect supervisor (again). I am also grateful to Miloš Kosterec, Igor Sedlár, Zsófia Zvolensky and conference participants for helpful comments. Thanks also to anonymous referees of *Organon F*.

References

ANGLBERGER, A. J. J., GRATZL, N. & ROY, O. (2015): Obligation, Free Choice, and the Logic of Weakest Permissions. *The Review of Symbolic Logic* 8(4), 807-827.

ANGLBERGER, A. J. J., FAROLDI, F. L. G. & KORBMACHER, J. (2016): An Exact Truthmaker Semantics for Permission and Obligation. In: Roy, O., Tamminga, A. & Willer, M. (eds.): *Deontic Logic and Normative Systems: 13th International Conference, DEON 2016, Bayreuth, Germany, July 18-21, 2016*. College Publications, Milton Keynes, 16-31.

BATENS, D. (2007): A Universal Logic Approach to Adaptive Logics. *Logica universalis* 1(1), 221-242.

BATENS, D. & HAESAERT, L. (2001): On Classical Adaptive Logics of Induction. *Logique et Analyse* 44(173-175), 255-290.

BEIRLAEN, M., STRAßER, Ch. & MEHEUS, J. (2013): An Inconsistency-Adaptive Deontic Logic for Normative Conflicts. *Journal of Philosophical Logic* 42(2), 1-31.

GOBLE, L. (2013a): Prima Facie Norms, Normative Conflicts, and Dilemmas. In: Gabbay, D. M. et al. (eds.): *Handbook of Deontic Logic and Normative Systems*. College publications.

GOBLE, L. (2013b): Deontic Logic (Adapted) for Normative Conflicts. *Logic Journal of IGPL Advance Access* 22(2), 8-29.

HANSSON, S. O. (2013): The Varieties of Permission. In: Gabbay, D. M., et al. (eds.): *Handbook of Deontic Logic and Normative Systems*. College publications.

HORTY, J. F. (1993): Deontic Logic as Founded on Nonmonotonic Logic. *Annals of Mathematics and Artificial Intelligence* 9(1-2), 69-91.

HORTY, J. F. (1997): Nonmonotonic Foundations for Deontic Logic. In: Nute, D. (ed.): *Defeasible Deontic Logic*. Springer Netherlands, 17-44.

CHEMLA, E., & BOTT, L. (2014): Processing Inferences at the Semantics/Pragmatics Frontier: Disjunctions and Free Choice. *Cognition* 130(3), 380-396.

KAMP, H. (1973): Free Choice Permission. *Proceedings of the Aristotelian Society*, 57-74.

KRATZER, A. & SHIMOYAMA, J. (2002): Indeterminate Pronouns: the View from Japanese. In: Yukio Otsu (ed.): *Proceeding of the 3rd Tokyo conference on psycholinguistics*, 1-25.

MULLINS, R. (2016): Rights in Default Logic. In: Roy, O., Tamminga, A. & Willer, M. (eds.): *Deontic Logic and Normative Systems: 13th International Conference, DEON 2016, Bayreuth, Germany, July 18-21, 2016*. College Publications, Milton Keynes, 187-202.

NOUWEN, R. (2018): Free Choice and Distribution over Disjunction: the Case of Free Choice Ability. *Semantics and Pragmatics* 11(4). https://doi.org/10.3765/sp.11.4.

REITER, R. (1980): A Logic for Default Reasoning. *Artificial intelligence* 13(1), 81-132.

SINGH, R., WEXLER, K., ASTLE-RAHIM, A., KAMAWAR, D. & FOX, D. (2016): Children Interpret Disjunction as Conjunction: Consequences for Theories of Implicature and Child Development. *Natural Language Semantics* 24(4), 305-352.

TIEU, L., ROMOLI, J., ZHOU, P. & CRAIN, S. (2016): Children's Knowledge of Free Choice Inferences and Scalar Implicatures. *Journal of Semantics* 33(2), 269-298.

VAN DE PUTTE, F., BEIRLAEN, M. & MEHEUS, J. (2018): Adaptive Deontic Logics: A Survey. Forthcoming in the *Journal of Applied Logics - IfCoLog Journal of Logics and their Applications*.

WILLER, M. (2017). Simplifying with Free Choice. *Topoi* 37(3), 1-14.

ZIMMERMANN, T. E. (2000): Free Choice Disjunction and Epistemic Possibility. *Natural Language Semantics* 8, 255-290.

Organon F 25 (4) 2018: 517-538

Verbeek on the Moral Agency of Artifacts

EHSAN ARZROOMCHILAR – DANIEL D. NOVOTNÝ[1]

ABSTRACT: One of the important questions discussed by philosophers of technology has to do with the moral significance of artifacts in human life. While many philosophers agree that artifacts do have moral significance attached to them, opinions vary as to how it is to be construed. In this paper we deal with the approach of the influential Dutch philosopher of technology Peter Paul Verbeek. He criticizes traditional ethical theories for assuming that whatever relevancy artifacts have for morality is entirely dependent on human beings, since artifacts are mere passive instruments of human agency. In contrast, he develops a view of moral agency that includes artifacts and that ascribes moral agency to human-technology hybrids rather than to humans as such. The goal of this paper is to elucidate Verbeek's account of moral agency and evaluate it. We also deal with his views on postphenomenology and mediation underlying this account. Although the general gist of our paper is expository, we point out to several problems for Verbeek's account.

KEYWORDS: Artifacts – mediation – moral agency – Peter Paul Verbeek – postphenomenology – technology.

[1] Received: 27 April 2018 / Accepted: 13 September 2018

✉ Ehsan Arzroomchilar
Faculty of Theology, University of South Bohemia in České Budějovice
Kněžská 8, 370 01 České Budějovice, Czech Republic
e-mail: ehsan.arzeroomchilar@gmail

✉ Daniel D. Novotný
Faculty of Theology, University of South Bohemia in České Budějovice
Kněžská 8, 370 01 České Budějovice, Czech Republic
e-mail: danieldnovotny@gmail.com

1. Introduction

Cars and trains, microwaves and refrigerators, TV sets and mobile phones, pencils, cups and glasses … countless artifacts from simple tools to sophisticated devices are ubiquitous in our lives. And there are many ontological, epistemological, ethical and other questions that philosophers may ask about these and other technological items. One of the main issues discussed by philosophers of technology today has to do with their moral significance. While many agree that technology and artifacts have moral significance attached to them, there are diverse views on how to construe it. In this paper we deal with the approach of the influential Dutch philosopher of technology Peter Paul Verbeek (University of Twente). In particular we deal with his account of moral agency, which is crucial for his project of reassessing the moral significance of artifacts.[2] Verbeek criticizes traditional ethical theories for wrongly assuming that whatever relevancy artifacts may have for morality is entirely dependent on human beings, since artifacts are mere passive instruments of human agency. In contrast, he develops a view of moral agency that includes artifacts and in which moral agency is ascribed to human-technology hybrids rather than to humans as such.

In what follows we first locate Verbeek's approach to the moral agency of artifacts in the broader context of contemporary philosophical studies related to technology (Section 2). Then we deal with postphenomenology, the philosophical background on which Verbeek draws (Section 3). Next we discuss one of his central concepts, namely that of mediation (Section 4), in order to better understand his view of moral agency (Section 5).

---

[2] We mostly draw on his *Moralizing Technology – Understanding and Designing the Morality of Things* (2011), where he develops his theory of moral agency most fully. We also take into account his other writings, especially his first book *What Things Do: Philosophical Reflections on Technology, Agency, and Design* (2005). We focus specifically on Verbeek's account of moral agency as we could not identify explicit discussion of agency in general. In order to facilitate a better understanding we occasionally provide direct references to authors that exerted great influence on him, namely Don Ihde and Bruno Latour. In doing this we by no means aspire to be exhaustive as our focus is Verbeek's theory as such, not an assessment of his reception of other authors. (We omit, for instance, references to Maurice Merleau-Ponty, Michel Foucault, Albert Borgman, Peter Sloterdijk and others.)

Finally, we evaluate his view and point out to some of its difficulties (Section 6).

2. Context

Verbeek's approach to moral agency of artifacts may be usefully related to two key questions that are commonly posed in philosophy of technology: (1) To what extent do humans shape technological products and processes? And (2) in what ways do technological products and processes shape human actions and experiences? (Mitcham & Waelbers, 2009).

As Mitcham and Waelbers have pointed out, in response to the first question we could imagine a spectrum wherein *voluntarism* is located at one end and *determinism* at the other. The advocates of voluntarism hold that the development of technologies is determined exclusively by human will, and that technological development is therefore malleable. Advocates of determinism, on the contrary, believe that technological development is determined by the internal logic of technologies themselves. Ellul, for instance, holds that old technologies are automatically replaced by those which are more efficient (Ellul, 1964; cf. Verbeek 2005, 11).

In response to the second question, we could propose another spectrum wherein *instrumentalism* is located at one extreme and *substantivism* at the other. Considering the relation between humans and technology, instrumentalists hold that technology is humanity's slave and thus it is nothing but an instrument in human hands. Advocates of substantivism, on the other hand, regard this relation as reverse and believe that technology is something "substantial", dominates over humanity and indeed holds us in its clutches.[3]

Now Verbeek's account of moral agency of artifacts is clearly neither voluntarist nor instrumentalist. Technologies are not determined exclusively by human will and they are not mere instruments. Relations between

[3] While these views are located at the two ends of a spectrum, adopting intermediary positions is also possible. For example, one can adopt a view in which neither humanity nor technology holds the other in its power. In such a perspective, the development of technology is due neither to human decisions alone, nor exclusively to the internal logic of technologies themselves. (Verbeek's view is a version of intermediary position).

humans and "things" are more complex. Already in *What Things Do: Philosophical Reflections on Technology, Agency, and Design*, first published in Dutch in 2000, Verbeek criticizes one-sided technophobia of earlier thinkers such as Heidegger and Jaspers that obscured intertwined, mediated character of human involvement with technology. He has been also increasingly interested in exploring the moral dimension of human-technology relations (Verbeek 2005, 212ff. and 2011). Drawing on earlier authors Verbeek approvingly acknowledges Hans Achterhuis's call for "moralization of devices" and Bruno Latour's view that "morality is not only to be found in humans but also in things" (Verbeek 2011, viii). As we shall see below, Verbeek provides various examples to show that "nonhuman entities are bursting with morality" (Verbeek 2011, 2).

Verbeek moreover vehemently opposes the "[m]ainstream ethical theory [that] does not leave much room for … a moral dimension of material objects" (Verbeek 2011, 2). What is the reason for this neglect? According to Verbeek it is the mistaken assumption that "technologies lack consciousness, rationality, freedom, and intentionality" and hence "morality ... is a solely human affair" (Verbeek 2011, 6). In order to amend this situation Verbeek sets out to develop "a notion of moral agency that does include material entities" (Verbeek 2011, 18).[4]

The above given quotes may seem to suggest that Verbeek sympathizes with determinism and substantivism in that artifacts are the main bearers of moral agency. However, although artifacts actively cooperate in shaping human experiences and actions we are not *completely* in their clutches. In fact, "ethics should be approached as a matter of human-technological association" (Verbeek 2011, 13). This means that "rather than separating or purifying 'humans and nonhumans' … the ethics of technology needs to hybridize them" (Verbeek 2011, 14). In other words, we cannot "hold on to the autonomy of the human subject as a prerequisite for moral agency; rather we need to replace the 'prime mover' status of the human subject with technologically mediated intentions" (Verbeek 2011, 16). In this way we get past the "subject-object distinction" and "articulate an 'amodern' perspective on ethics in which moral agency becomes a matter of human-

[4] Albeit Verbeek adds a cautious qualification here: moral agency of material entities "at the same time recognizes and articulates the differences between human and nonhuman elements of moral agency" (Verbeek 2011, 18).

technology hybrids rather than an exclusively human affair" (Verbeek 2011, 17).

We see that in his discussions of technology Verbeek refuses to separate humans from artifacts and hence his approach is best characterized as a peculiar intermediary position which is neither voluntarist nor determinist, neither instrumentalist nor substantivist. In order to understand Verbeek's approach to the moral agency of artifacts better let us now review what he says about "postphenomenology", the philosophical framework within which he develops his views.

3. Postphenomenology

What is postphenomenology? In the words of the American philosopher Don Ihde, its major proponent and initiator, it is "a modified, hybrid phenomenology" that avoids "the problems and misunderstandings of phenomenology as a subjectivist philosophy, sometimes taken as antiscientific" (Ihde 2009). Verbeek also understands it as "a new interpretation of the phenomenological tradition" but gives it "a broader definition" than Ihde (Verbeek 2005, 101). Verbeek's postphenomenology weaves together three streams: early philosophy of technology, phenomenology, and the empirical turn in technological studies from the 1980s (Verbeek 2005). Let us deal with them in turn.

3.1. Early philosophy of technology

Artifacts and technology have been the subject of occasional philosophical reflection since Plato and Aristotle (Schummer 2001). However, the urgency of the "problem of technology"[5] became obvious only in modern times. It was only in mid-nineteenth century Germany that sustained philosophical effort to understand technology led to the constitution of

[5] Peter Kwasniewski usefully characterizes the problem of technology in the context of his discussion of Leibniz as follows: "Leibniz's contribution to what may be called 'the problem of technology' (in the original sense of *techne* – art, craft, invention …) serves to highlight a major tension between belief in a fixed natural order providentially arranged for the best, and belief in a world of infinite possibilities, malleable to human hands and subject to human minds" (Kwasniewski 2017, 116).

a philosophical sub-field called "philosophy of technology" (Raydon) (Franssen et al. 2015). Most early contributors to the field lived and worked in Germany. Technology was a prominent theme in philosophers such as Karl Jaspers, Martin Heidegger, Hans Jonas, and others.[6]

For the most part Verbeek views the earlier, German phase of the philosophy of technology as a dialectics between Jaspers's existentialism and Heidegger's hermeneutics (Jonas's contribution is not explicitly taken into account). Whereas Jaspers was looking for answers to questions about the role that technology plays in human existence and action, Heidegger was searching for the way reality is disclosed by technology (Verbeek 2005, 16; Heidegger 1954/1977; Jaspers 1931/1951, 1958/1963). Verbeek takes the questions asked by Jaspers and Heidegger to be of crucial importance but finds their answers flawed. The most serious problem, Verbeek believes, is that they deal with the subject from a "transcendental perspective" (Verbeek 2005, 100). The transcendental perspective, as Verbeek understands it, is a perspective in which the researcher, instead of dealing with technological artifacts themselves and instead of examining their effects and consequences in daily life, addresses the "origins" of technology and the forces shaping it. For instance, in Jaspers's view technology is the fruit of a special functional way of looking at the world (Verbeek 2005, 28-30), while Heidegger holds that technology is the revealing of reality as a "standing-reserve of raw material" (Verbeek 2005, 95; Heidegger 1977, 10). Against this Verbeek calls for the study of technology which takes the particular artifacts themselves as its point of departure. This departure ought to issue in an evaluation of the outcomes engendered by those artifacts. We should not start with the artifacts and end with the causes and grounds of their formation. The direction of research should be exactly the reverse of the one found in the works of Jaspers and Heidegger.

In sum, what postphenomenology borrows from the early philosophy of technology are the questions: What kinds of impact does technology

---

[6] This earlier tradition of philosophy of technology, called now "humanities philosophy of technology" for its continuity with humanities and social sciences, has been complemented more recently by "analytical philosophy of technology", which is more closely related to philosophy of science and analytical philosophy (Franssen et al., 2015). For the most part Verbeek engages only the humanities philosophy of technology.

have on the existence and actions of humanity? How does technology affect the human experience of existence, and how does it disclose the world to us?

3.2. Phenomenology

The second major ingredient of postphenomenology is phenomenology. Phenomenology became established in the early twentieth century in the works of Edmund Husserl and authors influenced by him. It may be usefully characterized as "the study of structures of consciousness as experienced from the first-person point of view" (Smith 2013).

Phenomenology starts from phenomena as they appear in consciousness. In addition to phenomena, however, consciousness is also given and it is viewed as the place where humans and the world meet. The world is constructed/constituted in consciousness. And consciousness is always consciousness of something in the world. Therefore neither phenomena of the world nor consciousness can be found without the other. Human consciousness and the world can bear meaning only in relation with one another (Verbeek 2005, 109-112). It is this last point that is so attractive to Verbeek.

Verbeek and other postphenomenologists retain much sympathy for phenomenology. This is firstly because phenomenology's point of departure is to go back to the "things themselves". Indeed, postphenomenologists are interested in the study of technology from the artifacts themselves. Secondly, they agree with the phenomenological claim that the empirical sciences only represent *one* aspect of reality, not the fullness of reality as such. Nevertheless, despite acknowledging that our views of reality are always aspectual, phenomenologists assume that by employing their method they do get to represent "the true original world" in the end. Verbeek rejects this assumption (Verbeek 2005, 105). If phenomenology is to be used in studying technology, some modifications will be required. Verbeek makes the following two (Verbeek 2005, 104-116):

1. The possibility of any kind of access to the "original world" should be abandoned. Every sort of encounter with the world is "relative", not in the sense of an epistemological relativism, but rather in a more literal sense of the analysis of relations. It is the best approach one

can adopt, for "subject and object are not merely intertwined with each other but constitute each other" (Verbeek 2005, 112).

2. Phenomenology should be broadened to deal with any kind of relation between humanity and the world, including the relations that we are not conscious of. The relation between humanity and the world is not limited to the relations between conscious subjects and objects of which the subjects are conscious. The world is not just the source of *cognition* for humanity but, more significantly, it is where we *live*.

In sum, postphenomenology employs the phenomenological method, which is, however, modified in important ways. Most saliently it abandons aspirations to get to the "original world", emphases the interrelatedness of all experienced items and moves beyond what is presented in consciousness.

3.3. The empirical turn

The third main influence on Verbeek's postphenomenology has to do with a certain kind of empirical studies of technology emerging in the 1980s. Verbeek refers to Langdon Winner, the American scholar who discussed the low-hanging overpasses on Long Island in New York in "Do Artifacts Have Politics?" (Winner 1980). Those overpasses were deliberately built very low in order to prevent buses from using the road and allow only automobiles to pass underneath. At the time when these bridges were built this meant that racial minorities and the poor, who could not afford cars and who generally relied on public transportation, were effectively prevented from reaching the beaches. These overpasses shaped the ethnic and racial composition of people at the beach. What is remarkable about this investigation is its strongly empirical manner.

Another thinker acknowledged by Verbeek is the French anthropologist and philosopher Bruno Latour. He established a new framework by developing the so-called *actor-network* theory, studying artifacts as interactions between humans and artifacts. According to him humans and nonhumans are located in a network in a similar and indistinct way, with each component of that network cooperating. It is only the whole that acts. For example, in murdering someone with a gun, both the shooter and the gun are

responsible for that killing. This act is the result of the cooperation of the two "actants". Neither is able to carry out the task without the other (Verbeek 2005, 102). But while admiring the empirical character of such approaches to artifacts, Verbeek points out that we must not forget about the questions posed by the classical philosophy of technology (Verbeek 2005, 100).

Hence we see that Verbeek's version of postphenomenology pursues questions of earlier philosophy of technology by means of a modified phenomenology and a (certain version of) empirical studies pursued by Latour and others. Verbeek also stresses that postphenomenology should study both the hermeneutic and the existential aspects of artifact-human relations. It should deal with artifacts from the point of view of the role they might play in human perceptions of the world and how reality is thereby unfolded for humanity (the hermeneutic aspect). It should also deal with artifacts from the perspective of the role they might play in actions, behaviours, and in human existence generally (the existential aspect).

Having elucidated Verbeek's framework for studying technology we now turn to his views on mediation in order to interpret his account of the moral agency of artifacts.

4. Mediation

Artifacts are not just simple tools needed to attain human goals, their natures are not neutral. They have tremendous impact, sometimes foreseen and intended, at other times undesirable. Their influence in the world may even be unexpected and no one need be aware of it. Perhaps most importantly, however, artifacts at present *mediate* almost all our actions and perceptions. For instance, by sharing news and pictures of problems in developing countries, communication technologies have encouraged people in developed countries to spend more time and more money on charities (Waelbers 2011, 1). Another example: in some North European countries the length of the tube of the average vacuum cleaner is very short and thus is uncomfortable for men to work with, causing them back pain. This disinclines men to clean their homes and so reinforces sexist assumptions about who does housework (Waelbers 2011, 2).

Verbeek pursues the topic of mediation along the two lines of inquiry described in the previous section – hermeneutic and existential.

4.1. Hermeneutic aspects of mediation

Verbeek situates his account of the hermeneutic aspects of the mediation of artifacts into the context of the distinction between microperception and macroperception (Verbeek 2005, 122-123). The former concerns ordinary perception, such as seeing a tree or smelling a flower, while the latter concerns the framework within which sensory perceptions become meaningful. Artifacts have a huge impact on and hence a mediating role for both.

In their mediating role on the micro-level, artifacts lead to an amplification of some aspects and a reduction of others. For example, a thermographic camera shows some aspects of reality that we could never see without such equipment. At the same time some aspects of reality (including some non-visual dimensions) are reduced and distorted (Ihde 1991, 73-74).

Artifacts mediate human perceptions on the macro-level too. By changing the frameworks in which human interpretation occurs, they change our experiences. Postphenomenologists regard two of these frameworks as most important: cultural and scientific.

The cultural framework of macro-level mediation may be seen, for instance, in the appearance of communicative technologies and connections between different cultures. We are now prompted to see everything through different lenses, and so (typically) to show more tolerance toward different perspectives (Ihde 1993a). Also, thanks to the emerging modern technologies, humans are now compelled to make more choices; thus technologies create a "decisional burden" for modern humanity. This is obvious, for instance, in the advent of biomedical technology that forces people to make choices in situations that had traditionally been determined, such as when a foetus suffers from a serious disease detectable by our screening methods. We now have to choose between killing or letting "it" live. In Ihde's words, "The one choice I do not have is the choice not to make a choice" (Ihde 1990, 181).

The scientific framework of macro-level mediation concerns tools and equipment used in the constitution of scientific knowledge. This framework becomes ever more prominent as science plays an increasingly strong role in shaping the ways in which we interpret our world. We even evaluate our quotidian physical and mental well-being in medical and scientific terms. And it is also clear that scientific achievements are closely related to the advancement of our tools and equipment. Radio telescopes, for

instance, make things that are not accessible to the naked eye "perceivable". These mediated perceptions reveal entities that we would never have come to know without our mediating technologies. Technological instruments play an essential role in the generation of scientific knowledge, and studying this role is crucial for a proper understanding of contemporary science.

Artifacts, therefore, through altering human perception on micro- and macro-level play an undeniable role in how reality is revealed. Accordingly, postphenomenologists expand the notion of hermeneutics to apply it not only to texts but also to instruments and technological artifacts and their mediating roles. Hermeneutics in their hands becomes a kind of interpretation of objects (Ihde 1998, 139).

4.2. Existential aspects of mediation

Verbeek examines the mediation of artifacts from yet another angle, drawing on Latour's views. The core of his actor-network theory can be rendered as follows: Consider person A who murdered person B using a gun and aiming to take revenge. Latour's claim is that in such a situation we could not attribute this murdering only to the person who shoots the bullet, namely A; rather, the gun itself plays a mediating role in this event. In a Latourian perspective the scenario is this: A's "program of action" is taking revenge on B. On the other hand, the gun's program of action is shooting (not necessarily shooting a specific person). Out of these two programs of action a new third one arises, which is killing someone. This latter program of action is neither merely a result of A's program of action, nor exclusively a result of the gun's program of action; it is the outcome of a "composition" of both (Latour 1999). Latour sees all the actors, whether human or nonhuman, within a network in which they are constantly altering each other's program of action, resulting in a new program of action. In his words, "you are different with a gun in your hand; the gun is different with you holding it. You are another subject because you hold the gun; the gun is another object because it has entered into a relationship with you" (Latour 2005, 179-180). In his eyes, we should never see an artifact as a simple tool; rather we should consider it as having an agency analogically comparable to the agency of humans. "It will become more and more difficult to trace the border between the empire of the human and the realm of

technologies" (Latour 1992, 248). Artifacts thus continually shape our actions and deeds.

Latour's views are an important source for Verbeek who uses them to show that artifacts mediate both "how the world is present for human beings" (hermeneutic aspects of mediation) and how human beings are present in the world (existential aspects of mediation) (Verbeek 2005, 172). Verbeek also takes seriously Latour's thesis concerning the inseparability of humans and nonhumans.

5. Moral agency

Now, what do these points about postphenomenology and mediation have to do with the ethics and morality of artifacts? Since artifacts affect perceptions and human actions, they also affect morality. For instance, as we have seen, morality in the pathological form of racism became inbuilt into New York's overpasses. And Verbeek offers many other examples. Sonography is a method of examining foetuses by ultrasound. Consider a situation in which a pregnant woman finds, through sonography, that there is a high degree of probability that the child will be born with Down's syndrome. The finding may prompt the woman to abort the child and at any rate it will force her to decide what to do. We may also notice that the sonogram produces new meanings of what "foetus", "father", "mother" are and how they relate. It represents the foetus in specific ways: as being the same size as a newborn baby, in spite of the fact that it is much smaller, and as an independent entity freely floating in space, although it is closely linked to the mother. The sonogram isolates for us the experience of the foetus and separates it from its context, the mother (Boucher 2004, 12; Verbeek 2011, 24-25).

The sonogram also mediates the role of the mother and the father: the father, who formerly played an unimportant role, has now become a determining character in the new situation, deciding whether or not to abort. The mother, too, who formerly merely carried the baby, now, thanks to this instrument, has not only become a determining character but also a person whose uterus counts as posing a threat to the baby. On the other hand, depicting the foetus probably cements the bond between the mother, the father and the baby sooner than would happen without the technology and

may give the parents a feeling of being closer and more attached to the unborn child. In this way the mother and the father experience a new situation before the day of birth. But more than anything else, the aspect which makes the moral dimensions of this issue clearer is that with the sonogram the parents are converted from "expecting parents" to "deciding parents" (Verbeek 2011, 25-27). The sonogram brings about an unprecedented situation in which we could prevent the baby being born in the case of some kind of possible danger. The emergence of such dilemmas proves the ability of artifacts to create completely new moral situations. By mediating human perceptions and actions artifacts construct situations and objects, cooperating actively in depicting reality.

While the moral significance of artifacts is generally acknowledged, they are not commonly regarded as a substantial part of the moral sphere. Rather, they are seen as neutral and more or less unimportant tools (Smith 2003, 183). By contrast, according to Verbeek following Latour, the moral agency of artifacts is similar and comparable to that of humans (Waelbers 2001, 31-33). Verbeek, however, is aware of the difficulties facing those who make such claims. Moral agents should possess intentionality and freedom, they should be able to form intentions and realize them (Verbeek 2011, 54). Ordinarily, however, we do not think of artifacts as being intentional and free and hence Verbeek needs to explain to us in what way he thinks they are.

Let us begin with Verbeek's views on intentionality. We speak of intentionality in two senses – as the ability to form intentions, and as a kind of directedness (Verbeek 2011, 55). In Verbeek's view these two senses are related in a similar way as the hermeneutic and existential dimensions of the mediation of artifacts. The former concerns our perceptions of reality, whereas the latter our activities in reality (Verbeek 2011, 55-56). This means that our relations with the world are ordinarily mediated by artifacts. We either make contact with the world through artifacts, such as seeing the world through glasses, or technological artifacts shape our relation to reality as we make contact with the world, albeit remaining in the background, like the thermostats that automatically switch the heat on and off without our intervention. In all these kinds of relations, our intentionality is mediated by an artifact. Whenever we see beautiful scenery using binoculars, the intentionality is not just due to the human element, but seemingly due to a "human-artifact dyad". In other words, since the connection between

us and the world is not shaped just by our humanity alone, but rather with the help of artifacts, intentionality is not just a property of a human being but of a human-being-plus-artifact (Verbeek 2011, 56-58). As Verbeek puts it, "intentionality is distributed among the human and the nonhuman elements … [r]ather than being derived from human agents, [it] comes about in associations between humans and nonhumans" (Verbeek 2011, 58).

Now, Verbeek treats freedom in a similar way. Since artifacts don't have minds, it would seem inappropriate to ascribe freedom to them. However, artifacts often lead to completely unexpected outcomes as if they had their own minds. For example, energy-saving light bulbs were first used to decrease energy consumption, and are ordinarily cheap to run. As a consequence, people started to use those bulbs to light places that used to be dark (such as gardens), and eventually energy consumption increased. These and other examples lead Verbeek to think that artifacts should not be regarded as unfree. He also offers two more reasons to back up his claim. First, if we take freedom as an absolute concept, we could not count even humans as possessing it, since all people in all of their decisions are constrained by their era and the material environment and the artifacts to which they are related. So, to be free, it is sufficient to possess *some* degree of freedom (Verbeek 2011, 89). Such partial freedom is then also enjoyed by artifacts. Second, since human actions are mediated by artifacts, we should not think of human beings independently of their involvement with artifacts. Freedom is then a property of a human-artifact dyad. Hence, Verbeek concludes, freedom like intentionality is distributed among humans and artifacts (Verbeek 2011, 60).

We see that Verbeek has a way to ascribe intentionality and freedom to artifacts. Although, as we have seen, Verbeek appeals to various intuitions and considerations, his main reason for doing so has to do with his views of mediation and with his postphenomenological commitments. Human beings and artifacts co-constitute one another and form inseparable hybrids. It is these hybrids that are properly speaking intentional and free and hence moral agents. With respect to moral agency human beings are indistinguishable from artifacts. Both enter into the wholes that are moral agents in the proper sense, whereas taken alone they are not.

6. Problems

Within Verbeek's general postphenomenological commitments, as we have seen, it is not possible to ascribe moral agency to humans and at the same time to deny it to artifacts, despite Verbeek's occasional claims to the contrary. We take the general drift of his approach to blur any distinction between the moral agency of humans and artifacts.[7] In fact, Verbeek's postphenomenological understanding of (moral) agency does not seem to provide resources for drawing a clear distinction between humans and nonhumans – they are both parts of mutually constituted hybrids joined by mediation and other relations. He often insists on avoiding any kind of absolutizing subject and object (Verbeek 2005, 112). He also denies any gap between objectivity and subjectivity and speaks of mutual constitution. Artifacts (objects) and humans (subjects) are interwoven in such a way that they cannot be separated. In many passages Verbeek urges us to change our perspective on subjectivity and objectivity and, rather than assume them as pre-given, to consider them as co-shaped by one another (Verbeek 2005, 112). We, human beings, in some limited way do design and use artifacts, but they also structure our actions, perceptions and moralities. We stand in reciprocal relationships. We may initially decide to buy a car and use it, for instance, but immediately the car starts to affect our behaviour, expectations and thought. Once we have the car we may be able to rent a house far from our workplace, whereas without it we would have been obliged to live in the proximity of our workplace. Our behaviour has been affected by the fact that we own the car and as a result our situation within the world changes. There is no pure object vis-a-vis pure subject but all is

[7] At least at one occasion Verbeek claims that the idea that "technologies in themselves 'have' a form of agency that we normally only attribute to human beings" is a misreading of his work (Verbeek 2014, 79). He even notes that "it is in fact hard to find scholars who seriously defend the thesis that technologies can be full-blown moral agents just like human beings are" (Verbeek 2014, 79). We find these claims at odds with the general gist of his view. We hope to have made clear by now that Verbeek does not have resources to distinguish between the (moral) agency of artifacts and of humans. Also, by the way, it is not so rare to find scholars ascribing "full-blown agency" to some highly sophisticated artifacts such as AI robots, autonomous cars, etc. (These, however, are special subsets of artifacts, whereas Verbeek deals with artifacts in general).

"packed together" (Verbeek 2005, 164). The experiencing subject and the experienced object constitute one another.

Verbeek's claim that the subject is inseparable from the object allows him then to claim that "morality appears to be a coproduction of humans and nonhumans" (Verbeek 2014, 78) or that "morality is a hybrid affair" (Verbeek 2014, 80). One must overcome the view that morality is "located exclusively in humans" (Verbeek 2014, 80). The reason is simple – there is no pure human being, nor pure artifact.

The consequence of the human-artifact inseparability thesis is that human beings taken as such cannot be moral agents. Verbeek is aware that this calls for a new conception of moral agency. As he puts it: "rather than applying a human conception of agency to nonhumans, I rework the concept of agency in order to show that it should actually be seen as a property of hybrids rather than of humans only" (Verbeek 2009, 255). None of them could alone be deemed to be a self-standing agent. Morality is an attribute of a composite, of a network of human beings and artifacts.

There are three objections to Verbeek's view of the moral agency of artifacts that we would like to discuss.[8]

First, in our view Verbeek has misdescribed the moral status of artifacts by equalizing their contribution to moral acts. The conditions for an event to obtain ought not to be taken as a proper part of the event itself. Factors that bring about a specific framework within which a particular event happens are to be distinguished from the event itself. If I look at some beautiful scenery through a pair of binoculars, although this instrument does partly shape the framework of my experience, it is only *me* who is looking at that scenery, not me-*plus*-binoculars. The binoculars simply do not look at anything, they merely provide a framework within which I can see some things and not others. So even if it were appropriate to ascribe moral agency

[8] Other kinds of criticism have been put forward. Illis & Meijers (2014), for instance, object that Verbeek discusses only two necessary conditions of moral agency, intentionality and freedom, and ignores others. Philip Brey (2014) worries that by redefining moral agency and ascribing it to artifacts we are forced to ignore certain relevant features of human moral agents. Thorough and detailed criticism within the analytical tradition can be found in Peterson (2011) and (2017, 185); cf. also Selinger et al. (2012). While we are sympathetic to these kinds of criticism, our approach is more (although not exclusively) "internal", i.e. we point out to tensions within Verbeek's own philosophical commitments.

to artifact-human hybrids, it is humans rather than artifacts that are the primary locus of intentionality and freedom and hence of moral agency. The mediation of artifacts merely extends the sphere of moral agency which is grounded in and properly ascribed to human beings alone.

Why does Verbeek tend to obliterate distinctions between humans and artifacts? One of the reasons has to do with the way he describes his examples. True, no one had foreseen that the introduction of energy-saving bulbs would lead to an increase in energy consumption. This does not mean, however, that it was these light bulbs as such that decided that and hence are in the relevant sense responsible for it. We could have foreseen the danger and taken precautions. The light-bulbs could not. They are just what we make them to be. So, while it is true that artifacts dramatically change our lives and moralities and hence hardly are mere passive tools, they nevertheless do remain tools. It is to Verbeek's credit that he underscores the power of technology in our era and warns us about using and developing artifacts in an irresponsible way. However, we disagree with his account of the nature of artifacts and their moral agency. To highlight the role that technology can play in life one does not need to misrepresent the real functioning of artifacts.

Second, Verbeek's views on moral agency undo the distinction between artifacts and natural objects. If the only criterion that is at work in ascribing moral agency to things is whether it somehow affects the morality of actions, then (at least) some natural objects also qualify as moral co-agents. Hence we cannot distinguish them from artifacts. For it is clearly not just artifacts that structure our behaviours and steer our actions. Imagine, for instance, that Peter is walking in a dense forest and due to the existence of lots of trees and boughs he is obliged to constantly change direction. The trees and boughs act in the same way as a pair of binoculars does, except that they are natural objects, not artifacts. Does it make them moral agents as well? Is there any difference between the way that cars, knives or other artifacts affect our behaviour and that of the forest's effect? All of these put some specific restrictions on our activities, co-shaping our actions in a similar way. Or let's take another case. Suppose Mary runs into someone she hates and wants to take revenge on. Now imagine the following two possible scenarios. First, she takes a gun from her car and shoots the guy. He dies. Second, she leans over, picks up a big sharp stone and throws it at him. Again he dies. What is the difference? Both the stone and the gun

encouraged her to kill the guy and both shaped her actions. Stone-plus-Mary and gun-plus-Mary are both moral agents. Thus, Verbeek should acknowledge that (at least some) boughs and stones are moral co-agents. And if artifacts can be moral co-agents, then anything can. Perhaps Verbeek would be comfortable with this consequence. Many of us, however, would like to preserve the distinction between artifacts and natural objects and ascribe the status of moral co-agents only to some things.

Third, Verbeek has not left any place for the possibility of making a distinction between simple artifacts, such as a knife, and more evolved ones, such as autonomous cars. These are obviously not on the same level. For example, some of the more sophisticated artifacts may display abilities which make them more likely to qualify as moral agents than other simple ones. For a clearer grasp we can map out a spectrum representing various entities with respect to their intelligent behaviour dimension. In such a picture, we can locate natural objects at one extreme and human beings at the opposite one, with artifacts in between. It seems that not all artifacts could be situated at the same distance from humans. More complex artifacts, such as autonomous cars that need to "decide" how to react in unprecedented traffic situations, should be placed nearer to human beings than, for instance, knives. They imitate some aspects of human intelligent behaviour. Today's intelligent artifacts still lack some human abilities, such as moral deliberation or consciousness, but they do possess abilities such as learning, and (a sort of) thinking and decision-making. Perhaps eventually an AI robot will be constructed that will count as a full-blown moral agent. Simple artifacts such as flints or pencils, however, do not qualify. Our view, then, is that an adequate account of the morality of artifacts needs to do justice to the differences within their kind.[9]

Verbeek's remarks about the roles that artifacts can play in our lives are strikingly insightful. These observations should be taken seriously in designing and developing artifacts and in policy-making that concerns them. He has shed some light on how profoundly artifacts can change morality

[9] An anonymous referee points out that we assume in this objection that there is a hierarchy of the moral agency of artefacts, which may not be consistent with Verbeek's view about the "inseparability of humans and non-humans". We do not think so. And at any rate, even if our assumption is inconsistent is inconsistent with Verbeek's view, this only means that our objection is not internal but external.

and hence how important it is in applied and even in general ethics to take them into account. However, despite all of his contributions, the only lesson to take is that artifacts are much more powerful *tools* than we used to think, nothing less and nothing more. They are not as such agents nor co-agents, even though when we possess them there are lots of consequences for us humans. The ability of artifacts to change our lives requires us to become more careful and more responsible in developing and introducing technologies.

7. Conclusion

In this article we have dealt with Verbeek's view of the moral agency of artifacts. We have provided a broad philosophical background for his thinking by explaining the major elements of postphenomenology and the notion of mediation. We share Verbeek's sense of the urgency of "the problem of technology". A new technological invention usually profoundly modifies the moral situations that we have been facing so far. It is like "placing a drop of red dye into a beaker of clear water", to use Neil Postman's metaphor. After that we do not have clear water plus a spot of red dye but rather something entirely new (Postman 1998). Today ignoring the moral impact of artifacts is no longer an option. We value the contributions of Verbeek and other postphenomenologists to the ongoing debate about these issues. However, we have also found some aspects of his view, especially with respect to the moral agency of artifacts, wanting.

First, we have argued against placing artifacts and humans on the same level with respect to moral agency. In our view the Verbeekian approach by ascribing moral agency only to human-artefact hybrids runs the risk of anthropomorphizing artifacts and/or objectifying humans. Second, we have pointed out some undesirable consequences of Verbeek's views, namely the disappearance of the distinction between artifacts and natural objects. We think that philosophers of technology sensitive to phenomenological descriptions of our experiences should not abandon it. Third, we worry that Verbeek's claim that all human-artifact hybrids are moral agents hinders a proper understanding of the various levels at which some complex artifacts, such as robots or autonomous cars, may approximate moral agents while simple ones, such as binoculars or pens, do not.

We believe that Verbeek correctly shows that artifacts are not morally neutral, but in ascribing moral agency, albeit partial, to them he goes too far. His considerations undoubtedly show the need for responsibility in developing new technologies. However, technologies should not be assimilated to human moral agents. While artifacts profoundly affect morality, we cannot give up our own and proper individual responsibility as moral agents.

Acknowledgments

We would like to thank Martin Cajthaml, Světla Hanke Jarošová, Kateřina Kutarňová, Vojtěch Šimek, Peter Volek, participants of Bamberg-Budweis Philosophy Conference (Bamberg, July 4th, 2018), and especially to anonymous referees for discussion and comments. Novotný's work on this paper has been supported by Technology Agency of the Czech Republic (n. TL01000467 "Ethics of Autonomous Vehicles").

References

BOUCHER, J. (2004): Ultrasound – a Window to the Womb? Obstetric Ultrasound and the Abortion Rights Debate. *Journal of Medical Humanities* 25(1), 7-19.

BREY, P. (2014): From Moral Agents to Moral Factors: The Structural Ethics Approach. In: Kroes, P. & Verbeek, P. P. (eds.): *The Moral Status of Technical Artefacts*. Dordrecht: Springer.

ELLUL, J. (1964): *The Technological Society*. New York: Vintage Books.

FRANSSEN, M., LOKHORST, G.-J. & VAN DE POEL, I. (2015): Philosophy of Technology. In: Zalta, E. N. (ed.): *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), URL = <https://plato.stanford.edu/archives/fall2015/entries/technology/>.

HEIDEGGER, M. (1954/1977): *The Question Concerning Technology and other Essays*. Trans. by William Lovitt. New York: Harper and Row.

IHDE, D. (1990): *Technology and the Lifeworld*. The Indiana Series in the Philosophy of Technology. Bloomington: Indiana University Press.

IHDE, D. (1991): *Instrumental Realism*. The Indiana Series in the Philosophy of Technology. Bloomington: Indiana University Press.

IHDE, D. (1993a): *Philosophy of Technology: An Introduction*. New York: Paragon House.

IHDE, D. (1998): *Expanding Hermeneutics*. Evanston: Northwestern University Press.

IHDE, D. (2009): *Postphenomenology and Technoscience.* Albany, NY: SUNY Press.

ILLIES, M. (2014): Artifacts, Agency, and Action Schemes. In: Kroes, P. & Verbeek, P. P. (eds.): *The Moral Status of Technical Artefacts*. Dordrecht: Springer.

JASPERS, K. (1931/1951): *Man in the Modern Age.* Trans. by Paul Eden and Paul Cedar. London: Routledge & Kegan Paul.

JASPERS, K. (1958/1963). *The Atom Bomb and the Future of Man.* Trans by E. B. Aston. Chicago: University of Chicago Press.

KROES, P. & VERBEEK, P. P. (eds.) (2014): *The Moral Status of Technical Artefacts*. Dordrecht: Springer.

KWASNIEWSKI, P. A. (2017): Divine Wisdom, Natural Order, and Human Intervention. Leibniz on the Intersection of Theology, Teleology, and Technology. *Studia Neoaristotelica* 14 (2), 115-138.

LATOUR, B. (1987). *Science in Action: How to Follow Scientists and Engineers through Society*. Milton Keynes: Open University Press.

LATOUR, B. (1992): Where Are the Missing Masses? The Sociology of a Few Mundane Artefacts. In: *Shaping Technology, Building Society*. Cambridge, MA: MIT Press.

LATOUR, B. (1999): *Pandora's Hope*. Cambridge, MA: Harvard University Press.

LATOUR, B. (2005): *Reassembling the Social: An Introduction to Actor-Network Theory*. Oxford: Oxford University Press.

MITCHAM, C. & WAELBERS, K. (2009): Technology and Ethics: Overview. In: Berg Olsen, J., Pedersen, S. & Hendricks, V. (eds.): *A Companion to the Philosophy of Technology*. West Sussex: Wiley Blackwell, 367-383.

PETERSON, M. (2017): *The Ethics of Technology: A Geometric Analysis of Five Moral Principles*. Oxford: Oxford University Press.

PETERSON, M. & SPAHN, A. (2011): Can Technological Artefacts Be Moral Agents? *Science and Engineering Ethics* 17(3), 411-424.

POSTMAN, N. (1998): Five Things We Need to Know about Technological Change. Talk delivered in Denver Colorado March 28, 1998. URL = <http://web.cs.ucdavis.edu/~rogaway/classes/188/materials/postman.pdf>.

RAYDON, T. A. C.: Philosophy of Technology. *Internet Encyclopedia of Philosophy*, URL = <http://www.iep.utm.edu/technolo/>.

SCHMITT, R. (1967): Phenomenology. In: Edwards, P. (ed.): *Encyclopedia of Philosophy*. New York: Macmillan.

SCHUMMER, J. (2001): Aristotle on Technology and Nature. *Philosophia Naturalis* 38(1), 105-120.

SELINGER, E., IHDE, D., VAN DE POEL, I., PETERSON, M. & VERBEEK, P. P. (2012): Erratum to: Book Symposium on Peter Paul Verbeek's *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press, 2011. *Philosophy & Technology* 25(4), 605-631.

SMITH, A. (2003): Do You Believe in Ethics? Latour and Ihde in the Trenches of the Sciences Wars. In: Ihde, D. & Selinger, E. (eds.): *Chasing Technoscience: Matrix for Materiality*. Bloomington: Indiana University Press, 182-194.

SMITH, D. W. (2013): Phenomenology. In: Zalta, E. N. (ed.): *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), URL = <https://plato.stanford.edu/archives/sum2018/entries/phenomenology/>.

VERBEEK, P. P. (2005): *What Things Do: Philosophical Reflections on Technology, Agency and Design*. University Park, PA: Pennsylvania State University Press.

VERBEEK, P. P. (2009): Let's Make Things Better: A Reply to My Readers. *Human Studies* 32(2), 251-261.

VERBEEK, P. P. (2011): *Moralizing Technology: Understanding and Designing the Morality of Things*. London: University of Chicago Press.

WAELBERS, K. (2011): *Doing Good with Technologies: Taking Responsibility for the Social Role of Emerging Technologies*. New York: Springer.

WINNER, L. (1980): Do Artifacts Have Politics? *Daedalus* 109(1), 121-136.

Organon F 25 (4) 2018: 539-563

Theories of Reference and Linguistic Relativity

ANTONIO BLANCO SALGUEIRO[1]

ABSTRACT: The challenge to traditional theories of reference posed by experimental philosophers puts the focus on the question of diversity, cultural and linguistic, on the one hand, and cognitive (on intuitions), on the other. This allows for a connection between the problem of reference and the language-thought relation debate, and the linguistic relativity hypothesis conceived as the idea that linguistic diversity causes a correlative cognitive diversity. It is argued that the Kripkean view on proper names and natural kind terms is probably universal and that this empirical fact has plausible consequences for the universality of certain forms of human thought, but that there are nontrivial differences in the details of the workings of these expressions in different languages and that those differences influence the ways of thinking of speakers about individuals and kinds.

KEYWORDS: Language – linguistic diversity – linguistic relativity – names – reference – thought.

1. Introduction

The motivation underlying this work is the challenge posed by experimental philosophy to the theories of reference in Machery, Mallon, Nichols

[1] Received: 20 April 2018 / Accepted: 14 September 2018

✉ Antonio Blanco Salgueiro

Department of Logic and Theoretical Philosophy, Faculty of Philosophy
Complutense University of Madrid, 28040 Madrid, Spain

e-mail: ablancos@filos.ucm.es

& Stich (2004) and a long series of further articles.[2] They argue that the intuitions invoked by philosophers (according to those authors, as the sole evidence for their views) aren't universal, but vary across cultures and, in particular, differ substantially when comparing people from the West with people from the Far East.[3] Their experiments appear to show that when presented with stories such as Kripke's Gödel-case, Chinese participants tend to have descriptivist intuitions, while Americans tend to have Kripkean intuitions (Mallon *et al.* 2009, 34). The data also appear to show that the diversity is not only cross-cultural but also intra-cultural: 45% of Americans gave descriptivist answers (as did most Chinese), and 30% of Chinese gave Kripkean answers (as did most Americans). The subsequent work of this group of philosophers respond to a huge critical literature. The controversy doesn't only reach philosophy of language. The point is metaphilosophical, on the role of intuitions in the philosophical endeavor at large, and on the very idea of *intuition*.[4] Their aim is to question the project of constructing a theory of reference or any other philosophical theory (in fields like ethics, epistemology, etc.) taking as evidential ground the intuitions of lay people (or of experts). Some of the criticisms to Machery *et al.* seem fair to me, and I shall not enter into some of the more heated debates. In my view, the crucial point in their challenge is to place philosophy of language's focus on the problem of diversity. Firstly, on linguistic diversity, and secondly, on the possibility of a correlative cognitive diversity (after all, intuitions are mental states of speakers). This is what allows us to connect the controversy about experimental philosophy with the question of linguistic relativity, conceived as the idea that linguistic diversity (differences in the ways of speaking) brings along a correlative cognitive

---

[2] Some works of experimental philosophers are Weinberg *et al.* (2001), Machery *et al.* (2004), Mallon *et al.* (2009), Machery *et al.* (2009), Machery *et al.* (2010); Machery (2012); Machery *et al.* (2013), Machery *et al.* (2015), Machery, Sytsma & Deutsch (2015), Nichols *et al.* (2016), Stich & Tobia (2016).

[3] Mallon *et al*. (2004) rest on the ideas of cultural psychology (Nisbett 2003; Nisbett *et al.* 2003).

[4] This is just one sample of this critical literature, Liao (2008); Deutsch (2009); Jylkkä *et al.* (2009), Martí (2009); Lam (2010); Ludwig (2010); Ichikawa *et al.* (2011); Devitt (2011, 2012); Knobe *et al.* (2012); Vaesen *et al.* (2013); Sytsma *et al.* (2015); Nado & Johnson (2016); Heck (2017); Hannon (2017).

diversity (differences in the ways of thinking). The main idea that I intend to explore is that the differences in the mechanisms for the reference of certain types of expressions create differences in the ways in which humans mentally refer to individuals and kinds.

In Section 2, I highlight the empirical character of the claim that the mechanisms for reference are universal. Then, I present two languages in which names function in a different way from English (Section 3). The claim is not that they work in a non-Kripkean way, but that there are differences in the details of the mechanisms involved. In Section 4, I present the argument for linguistic relativity. Based on its first premise, in Section 5, I argue for a link between the universality of the Kripkean character of names and the universality of counterfactual thought, which amounts to a semiotic effect of *any* language on human cognition. In Section 6, I distinguish different kinds of linguistic diversity, and in Section 7 I connect them with possible differences in the referential mechanisms that could in turn cause differences in referential thought. Some implications are finally drawn from this hypothetical impact of language diversity on thought diversity.

2. Kripke and the universality of the mechanisms of reference

Does Kripke offer an account of proper names which claims universal validity, i.e, that intends to explain how proper names work in all human languages, real or possible, past, present or future? Of course, the same question may be addressed to an advocate of any other theory of names, in particular to any type of descriptivist, or to hybrid views. Most of Kripke's readers will answer affirmatively and assume that Kripke and they themselves aren't just studying how proper names work in English, although he uses only English examples and is not very explicit on this point.

Many linguists that have renewed the interest in linguistic diversity (Crystal 2000; Evans & Levinson 2009) highlight that throughout the decades of full Chomskyan hegemony, it was normal for someone to have a successful career in linguistics without studying or being fluent in any language except English, let alone in (what for us is) an exotic language. This holds still more accurately in the field of philosophy of language, where anything that aspires to have some relevance should be written in English.

But are we sure that we are doing philosophy of language and not philosophy of English?

One option would be to ground the answer in human biology. Are we all born with a disposition to use names in a causal-historical way? This nativist Kripkeanism could apply the "argument from the poverty of stimulus" and argue that without enough evidence children would begin to use proper names in a Kripkean way, not in a descriptivist way. This could be true, although I think it is not so.[5] Still, we would need an evolutionary account of why humans acquired this predisposition in the phylogenesis (that manifests itself in the ontogenesis), that is, what advantages our Kripkean ancestors had over their descriptivist rivals that allowed them to reproduce more profusely. Or this may be a brute fact that occurred by chance: humans could have been descriptivists, but they (or most of them, anyhow) just happen to be Kripkeans. In any case, Kripke never mentions human biology. In the Preface to *Naming and Necessity* he says that proper names are rigid designators *de jure* because their reference "is *stipulated* to be a single object, whether we are speaking of the actual world or of a counterfactual situation" (Kripke 1972/1980, 21). The question is what guarantees that in every language this stipulation for the use of proper names is made? Talking of "stipulation" seems to imply that things could have been different: we could decide not to stipulate this and stipulate another thing instead. If this is a universal rule present in every one of the over 6000 languages now spoken around the world and in the much larger number that have ever existed, there must be some universal pressure that accounts for it, and the search for it should be a central concern for a theory of reference. Most philosophers agree that the reference of proper names is not a biological question, but depends on conventions, rules, practices or language games. However, we find diversity in other fields with these same features. Why not in the rules that establish how proper names refer? Just before the famous passage where he exposes his new picture, Kripke admits that we are free to stipulate that our names work as the descriptivist says that they in fact work. There is then no human (let alone physical,

[5] I don't mean to deny that children are born with a pre-linguistic capacity for individuating objects, but only that this capacity determines by itself the correct theory of reference for every human language.

metaphysical or logical) necessity that guarantees that all humans will stipulate that names work causal-historically:

> So what makes my use of 'Cicero' into a name of *him*? The picture which leads to the cluster-of-descriptions theory is something like this: One is isolated in a room; the entire community of other speakers, everything else, could disappear; and one determines the reference for himself by saying—'By "Gödel" I shall mean the man, whoever he is, who probed the incompleteness of arithmetic'. **Now you can do this if you want to. There's nothing really preventing it. You can just stick to that determination.** If that's what you do, then if Schmidt discovered the incompleteness of arithmetic you *do* refer to him when you say 'Gödel did such and such'. (Kripke 1972/1980, 91; my boldface added)

Here Kripke is clear that it is possible in principle to create names that work in the way the descriptivists think they work in English. His claim is that this is not how we use them as a matter of fact and that there are good reasons for using them as we do. He suggests that only a weird speaker would use his names this way, in a sort of private language. This would open the door to an individualistic descriptivism where a speaker applies her own descriptions without taking into account the descriptions of others. But what prevents a whole community from creating a Descriptiranto in which names work descriptively for every speaker? And what allows us to discard that in some actual languages, with no deliberate decision, but in the tacit way in which many conventions are established, names work descriptively? In fact, the descriptivist philosopher thinks that English is such a language; we can reverse the question: What prevents proper names from working causal-historically in some languages? I take this as being the basic challenge that arises from the controversy initiated by Machery *et al.* (2004).

My own answer is that it is possible that the existence of proper names that work causal-historically is a linguistic universal, but that we cannot take it for granted and that the question has to be decided empirically. Moreover, I believe that Kripke's view is right for most names in English, but it is good to remember that there are English-speaking philosophers that are descriptivists concerning the functioning of proper names in their own language. This shows the magnitude of the problem: if it is difficult to settle the question for languages in which scholars are fluent, the difficulty

can be higher for non-familiar languages, not to mention for human language in general.

Linguist Daniel Everett offers an analogy that illustrates the kind of functional account that could be alleged for the features that are found in all, most or many, languages: the analogy with the independent invention of the bow and arrow in many different parts of the world. It seems absurd to postulate an innate faculty for making bows. But there exists a general pressure, killing protein that moves faster than we do (Everett 2013, 17), which explains why smart beings like humans find the same solution again and again. We find an account of this type in the classic Putnam (1975). His first argument against descriptivist theories of natural kind terms is based on a universal linguistic fact (the division of linguistic labor) that depends, in turn, on a hypothetically universal human practice (the division of non-linguistic labor).

Kripke suggests that his picture follows from very general facts about humans and their relationship with language. In fact, Kripke seems to consider that it follows from something as general as the fact that language is a social more than an individual tool (Kripke 1972/1980, 163). Should this be right, we could refer the universality of the causal-historical theory to the universality of language as a social tool. After all, the problem of killing protein that moves faster than we do is not more obviously universal than some of the problems whose solution is alleged to involve a use of names according to the causal-historical model, like the problem of talking about non familiar persons or places with respect to which some or most of the members of the community can have false or non individuative beliefs, or the problem of talking about what could have happened to a person, different from what really happened to her, that is, the problem of considering counterfactual scenarios about particulars.

An important non-empirical part of Kripke's work belongs to what can be called *philosophical linguistic typology*; it differs from the typology of linguists in its interest in the workings of certain types of expression in *possible* languages. So, it is not directly conditioned by the empirical findings that could be alleged as a result of the study of specific natural languages. We don't need to visit the Amazonia to do this or even to give functional reasons related to the use of certain expressions in some human practices. We can do it from our comfortable chairs of Western philosophers, at the risk of not including possibilities that in fact occur in actual languages of the

world. It is at least remotely possible that philosophers don't have the extraordinary imaginative capacities that they so often assume to have, and that allows them to visit every corner of the logical space without leaving their offices. Kripke's typology of designators belongs to this part, and the very definition of "designator" as a wide category with several subcategories. He is clear in establishing that in this first part he is neutral with respect to what types of designator actually exist in a particular language like English, or with respect to the hypothesis that we will find the same classes in every human language, i.e., that some types of designator are *semantic universals*. As is well known, his main distinction here is between *rigid* and *accidental* designators, with a subdivision of the first type in rigid designators *de jure* and *de facto*. We could also include here the distinction between *semantic reference* (the one conventionally associated with the expressions of a language) and *speaker's reference* (linked to speaker's intentions, independently of the conventional use of the expressions he uses) (*cf.* Kripke 1977).

The second moment is *empirical* and much more controversial. Kripke advances an empirical thesis about proper names and other classes of expressions as we find them in English and perhaps in any natural language, although he admits the possibility of inventing artificial languages without rigid designators. It is here where the controversy arises over whether the intuitions of competent speakers are the only source of evidence that can decide the question of semantic diversity. Machery *et al.* claim that philosophers allege only those intuitions, while Devitt and others claim that the main source of evidence for a theory of reference does not come from intuitions (be they from lay people or from philosophers of language, that he thinks are bound to be better than those of the common folk), but from the overt (spontaneous or elicited) *use* of names by competent speakers (*cf.* Devitt 2011; 2012), a stance that I basically agree with. The priority would then be to discover if all humans use proper names in the same way.

Another empirical point is the view of how a proper name or other rigid designator is connected with its bearer. Kripke claims that descriptivism derives its plausibility (apart from solving some puzzles) from the fact that it proposes a *mechanism* that removes the apparent magical character of this link.[6] If we grant that Kripke offers an alternative mechanism for reference,

[6] There can be a lack of harmony between the mechanisms that really do the work and the intuitions of the speakers, if they are biased by cultural myths about language.

then we can see a way to connect the typological and the empirical parts (hybrid theories propose combinations of both mechanisms). Some formal types of designator may not have plausible mechanisms that could realize them in the real world. It is possible that a language for supernatural beings be not constrained by facts about the natural and social world, but a language for humans clearly is. One thing is that we can invent languages which contain by stipulation one, several or all types of designator, and another is whether these languages could work or be used by beings like us in a world like ours. From Kripke's work on, it is assumed that there are different mechanisms that could work in our physical world to back the reference. It is no longer the case that the descriptivist wins because otherwise reference would be mysterious. In fact, for Kripke there are in English both expressions that work causal-historically and expressions that work descriptively (most definite descriptions).

3. Two "exotic" languages: denk nicht, sondern schau! [Don't think, but look!]

I have suggested that there may be some cross-linguistic practices that can justify the universality claim of a theory of reference (in particular, that of Kripke's) for proper names and other kinds of expressions, such as the practice of talking of individuals or substances about which one has insufficient or erroneous knowledge, the practice of ascribing mental states about individuals or substances to a person with such a defective knowledge, or the practice of considering modal situations. However, we shouldn't take for granted that names are used in the same way everywhere, or that cultural practices don't have any impact on the semantics of this class of terms. I shall offer two examples taken from field linguistics. Unlike philosophers, linguists pay little attention to proper names or to the

This is one of the criticisms made to Machery *et al.*'s position (Martí 2009). Think of a cultural belief in the magical powers of names: you can influence someone through her name. For these people it would be natural to claim that there is a magical bond between a name and its bearer, beyond any description or something as prosaic as a causal-historical chain. But surely the more plausible real bond is one of those proposed by naturalist theories of reference.

problem of reference, but some of the things they say have obvious consequences for the theory of reference. Both examples show that proper names are used in other cultures in quite different ways from ours and that it is risky to claim that those differences will never affect the heart of our preferred theory of reference (there are probably more radical cases). I think that these kind of cases shed more light on the theory of reference (taken as a form of "experimental semantics") than the method of consulting the intuitions, which are not even mentioned in these studies (although it would be interesting to test the intuitions of speakers of these languages).

3.1. Jesus' name

The first case is taken from a study of the Amazonian language Pirahã (Everett 2008). Everett's controversial thesis (firstly proposed in Everett 2005) is that a cultural principle permeates the form of life of the Pirahãs and is responsible for many features in the grammar of their language, the *principle of immediacy of experience* (PIE), according to which the communication is limited to the immediate experience of speakers. The idea of immediacy doesn't imply something so radical as to stick to the present moment. An experience is immediate in Pirahã if it has been seen or recounted as seen by a person at the time of telling (Everett 2005, 622).[7] The argument for taking culture as the causal factor is that there are many rare or unique features of the Pirahã language that are formally very different but that can be connected and explained by this unique cultural principle.[8] The claim is that the Pirahã culture has a holistic impact on their language, showing up in various aspects of it as a coherent way of speaking in accord with the aforementioned principle. Everett doesn't say much about the use of proper names in Pirahã, but clearly the PIE affects them, given that it prevents talking about particulars with respect to which no member of the

[7] This cultural principle implies a restriction to the epistemology of testimony. Epistemology is important for the philosophy of reference. One of the disputed questions is whether we must know the referent well enough for a name to refer to it. Far from thinking that Pirahã's epistemology is defective, Everett argues that it is better than ours (which is, of course, questionable).

[8] The most controversial idea is that Pirahã lacks recursion altogether, but Everett also points to other shocking absences (and some presences) such as the lack of numbers, quantifiers or fixed color terms.

community has had a direct familiarity. For this reason, one class of names that responds well to the Kripkean view, the names of historical characters, doesn't exist in Pirahã. In fact, they don't have creation myths or a mythical or real history. Everett's life with the Pirahã people began as a project of evangelization aimed at converting them to Christianity, but he himself was converted to atheism, after not being able to convert a single Pirahã.[9] What is interesting for us is his failure to introduce the name "Jesus" in the naming practices of the Pirahã (they don't have a name for God either, and he used *Baíxi Hioóxio*, "Up-high father"). Many Westerners think that "Jesus" is like "Jonas", the name of a real but legendary man who wasn't born from a virgin, didn't work miracles, didn't resurrect, etc. When the Pirahã asked if he himself had seen Jesus and he tried to explain that Jesus had lived a long time ago, he couldn't get them to understand him or take him seriously. All the names of persons in this culture are names of live people or of someone closely known by a live speaker: "the Pirahãs believe only what they see. Sometimes they also believe in things that someone else has told them, so long as that person has personally witnessed what he or she is reporting" (Everett 2008, 266).

I don't want to claim that proper names don't work causal-historically among the Pirahã. That depends on the existence of practices like speaking of somebody using her name in cases of ignorance or error (which is less probable given their cultural emphasis on the evidence), ascribing mental states in these circumstances, or considering counterfactual scenarios (I assume that these arguments for Kripkeanism are correct); the PIE could also prevent all this, but I ignore it. What the case shows is that there can be restrictions to the causal-historical links between a name and its bearer, that is, to the intention of fluent speakers of making the reference of some types of expression depend on these links. What is more, proper names seem to work among the Pirahã in the same way as Strawson (1959, 82) describes our own use: some speakers, with respect to some names, can "pass the buck" to others who are supposed to know better, but there is always at

[9] The Summer Institute of Linguistics entrusted Everett with the study of Pirahã to translate the Bible into this language. According to him, "missionaries had been trying to convert them for over two hundred years", but "no Pirahãs are known to have 'converted' at any period of their history" (Everett 2008, 269).

least one member of the community that can give an accurate and individuative description of the referent.[10] In any case, part of what Everett says about the Pirahã, such as that regarding the complete absence of fiction or myths among them, point to severe restrictions to counterfactual discourse in this community, which can undermine the universal application of the "modal argument" for the causal-historical view on proper names. The question as to whether Pirahã names work descriptively or causal-historically (with strict epistemological restrictions) requires a more careful empirical scrutiny.

3.2. Naming in Madagascar

We travel from the Amazonia to Madagascar and from Pirahã to Malagasy, an Austronesian language with 18 million native speakers in 2007. My source now is Och (1976), a field work on the conversational practices in traditional communities in Madagascar. I shall consider only what she says about names, but her general aim is to question the universality of Grice's conversational maxims. Again, for cultural reasons, some of these practices affect the use of names for persons. The natives think that if someone is called by his name, the spirits can cause him harm through it. For this and other cultural reasons, they usually avoid the reference by the name, letting it remain implicit, or use descriptions as the equivalent of "the builder" or even general descriptions like "the person", so that it is normal for a mother to ask her son, referring to her husband: "Has the person already come?" For the same reason, they change their names when they suffer a misfortune. In recent times the authorities have forbidden

[10] An anonymous reviewer objects that the difference between English and Pirahã could be "a difference in the standards of testimonial justification rather than in the language". But it should be borne in mind that according to Everett the PIE has a holistic impact on the Pirahã language, affecting many aspects of its structure. So, this would not be a minor epistemic restriction on some autonomous linguistic mechanisms. This culturally entrenched epistemology is supposed to be widely codified in the language and constantly reinforced through its use. All I am adding here is that if Everett is right, the PIE also affects the rules for proper names. More generally, Everett's view is akin to the Wittgensteinian idea of the intertwining of language and life forms, against the idea of the autonomy of language with respect to culture (Everett 2005, 622). In Wittgenstein's terms, "Jesus" doesn't find a place in the language games of the Pirahã.

changing one's name more than three times in the course of a life, because in the past it was usual to change it six or seven times.

As before, we can ask what the implications are of all this for a theory of reference with universalist ambitions; at least, the functioning of proper names among these people is not exactly like ours, as the causal-historical chains that can be invoked are shorter and more dubious than the ones associated with our proper names. The feeling that there is no profound difference here with our linguistic practices stems from the assumption of an ontology of persons according to which a change of name does not imply a change of referent in these cases. But in many instances of name change in languages like this, the assumption is that the referent is not the same or exactly the same after the change; that is, there can be ontological differences associated to this linguistic diversity in the naming practices. To appreciate this, we can resort to some exceptional similar examples among us, like the "Cassius Clay" / "Muhammad Ali" case. A person who changes her name after her religious conversion can say very seriously that she is not the same person as before. The apparently trivial application of the law of identity in "Cassius Clay is Muhammad Ali" would be problematic if there were social consensus that the individuals are different. At the very least, in many cultures a change of name is linked to a change of social status, in the social identity of the individual. Here, language performatively creates new social reality (new social persons).[11] This idea that linguistic diversity implies diversity in the kinds of social reality that language

[11] The inconstancy in the use of proper names is present in other languages. In fact, we find it in Pirahã. Everett claims that once, after a prolonged absence, he addressed a pirahã using what he thought was still his name, and that the pirahã did not react. The following quote also illustrates other (for us) oddities in the institution of names among this people:

> One of the men, Kaaboogí, […] addressed me in very rudimentary Portuguese: "*Pirahã chamar você Xoogiái*" (The Pirahãs will call you OO-gi-Ai). I had received my Pirahã name.
>
> I knew that the Pirahãs would name me, because […] they name all foreigners, since they don't like to say foreign names. I later learned that the names are based on a similarity that the Pirahãs perceive between the foreigner and some Pirahã. Among the men there that day was a young man named Xoogiái, and I had to admit that I could see some resemblance. Xoogiái would be my name for the next ten years, until the very same Kaaboogí, now called Xahóápati, told me that my name

can create is one of the "new directions" in the study of linguistic relativity (Enfield 2015, 216). Again, even if the general mechanism for reference were causal-historical (allowing, for example, a certain type of modal discourse about individuals), there can be nontrivial differences across languages in the details of the implementation of this common mechanism, connected to different functions of words in cultural practices.

4. Linguistic relativity

Linguistic relativity is often defined (by its detractors) in ways that make it seem a radical and implausible idea. I take it simply as the claim that linguistic diversity (the different realizations of human language) has a nontrivial impact on cognitive diversity (the various styles of thinking in humans). In Blanco Salgueiro (2017) I provide a map of the many forms that this hypothesis may take, assuming that it is the conclusion of an argument whose two premises can be formulated in different ways. The radicalism or moderation of the hypothesis is the result of adopting one or other version of the premises. This is my reconstruction of the general argument, with many of the complexities in brackets:

Premise 1 *Cognitive Impact of Language* (CIL): Language [such and such aspects, features, levels or mechanisms of any particular language] has [strong or weak, qualitative or quantitative] effects on thought [in such and such cognitive domains; in the most classic version, on the interpretation or construction of reality] and on behavior.

Premise 2 *Linguistic Diversity* (LD): But the different languages [or linguistic variants] differ among them [little or much] in [some or all] the aspects that cause those cognitive or behavioral effects.

---

was now too old and that my new name was Xaíbigaí. (About six years after that my name was changed again to what it is today, Paóxaisi – the name of a very old man). As I learned, the Pirahãs change names from time to time, usually when individual Pirahãs trade names with spirits they encounter in the jungle. (Everett 2008, 9).

Conclusion *Linguistic Relativity* (LR): So, there exist cognitive and behavioral differences between speakers of different languages [or linguistic variants].

I shall not argue here for the plausibility of some of the argument's versions; my aim is just to connect it with the problem of reference. But three related points must be highlighted. Firstly, most current relativists reject deterministic versions of Premise 1 ("linguistic determinism"), and argue for a weaker but nontrivial *influence* of language on thought ("linguistic influencism"). Secondly, although often the question discussed is the relative priority of *language* and *thought*, probably the relationship is dynamic: not language affecting thought (L→T), or thought affecting language (T→L), but both interacting in complex ways (L↔T). Moreover, further factors, like culture, could play a key role. This last possibility has gained strength in linguistics, partly thanks to Everett's work on the Pirahã language which, in his view, defies for cultural reasons the most basic ideas about Universal Grammar (like the universality of recursion). But in his first controversial work, Everett explicitly rejects LR:

> […] against the simple Whorfian idea that linguistic relativity or determinism alone can account for the facts under consideration. In fact, I also argue that the unidirectionality inherent in linguistic relativity offers an insufficient tool for language-cognition connections more generally in that it falls to recognize the fundamental role of culture in shaping language. (Everett 2005, 623)

To this, a neo-Whorfian responds what will guide my following remarks:

> […] a language of course is a crucial part of a culture and it is adapted to the rest of it […] The question that neo-Whorfians are interested in is how culture gets into the head, so to speak, and here language appears to play a crucial role: it is learnt far earlier than most aspects of culture, is the most highly practiced set of cultural skills, and is a representation system that is at once public and private, cultural and mental. It is hard to explain nonecologically induced uniformities in cognitive style without invoking language as a causal factor. (Levinson 2005, 638)

Previously, in the Introduction to a classic in the neo-Whorfian literature (Gumperz & Levinson 1996, 1), the editors set a link between culture, language and thought in their very definition of LR as the idea that culture, *through* language, affects the way we think. I take this as the orthodox stance in modern defenses of LR. This is close to what Whorf himself claims in Whorf (1939), where he gives diachronic priority to the culture (he sees language as a cultural construction), but also insists that language is the strongest factor synchronically, accounting for how individual thought is affected by social factors. In non-biological approaches to language the distinction between language and culture is not neat.

Machery *et al.* (2004) assume without question that the key factor that explains the differences in intuitions is culture, not language. They are so sure that they don't see a problem in conducting their experiments in English, when comparing American English speakers and Hong Kong speakers, whose mother tongue is Cantonese, for whom English is a second language, and who could make transfers from their first to their second language, a well-known phenomenon in the study of second languages ("false friends"); so it is unclear that the differences are due to culture and not to language. Lam (2009) criticizes this part of their methodology. In reaction, Machery *et al.* (2010) repeat the experiments using the native tongue for each group; for the Chinese group the stories appear now in Chinese writing, common to Cantonese and Mandarin. They claim that the results are analogous to those of the original experiment. In my view, the most plausible hypothesis (following Levinson's argument) is that language diversity is the direct cause of the cognitive differences in this case, if they are confirmed, although the cultural forms of life may be the ultimate causes. If there were differences in the referential conventions associated to the designators of different languages, and taking into account the early acquisition and habitual character of the practices that involve the use of names, it is plausible that this has some impact on the differences in cognitive style, i.e., that we habitually think using the same conventionally established referential mechanisms that we use when we speak.

5. From linguistic reference to mental aboutness

But we shouldn't rush to accept that we will find diversity in this field. If it were true that there is a universal pressure that guarantees the universality of the mechanisms for reference, then the possible impact of language on thought could also be universal. That is, as in other domains, Premise 1 (CIL) of the argument for LR can be held independently of the truth of Premise 2 (LD). Many authors have claimed that the most important impact of language on thought is transversal to languages. It would be the fact that we are verbal beings and not the fact that we speak a particular language, which explains the human cognitive singularity (our capacity for planning, regulating our actions, thinking about thinking, nonmodular thinking, active thinking, etc.).[12]

Let's assume that the Kripkean view is correct and that this follows from our need to invoke it to explain how speakers talk counterfactually (they keep applying a name to a particular even when they imagine that the descriptions associated to it were false); or to explain how they refer to a particular using a name in cases of ignorance or error (and of semantic reference, not of speaker's reference); etc. Then, we can formulate a special case of Premise 1, the claim that human language is what allows or at least fosters or facilitates counterfactual thought, or thought about particulars in cases of ignorance and error in humans. The hypothesis predicts that nonverbal beings don't have, or are less good at, those kinds of thought.[13] The mechanism involved could be the same as the one invoked in much current research on linguistic relativity,[14] namely, *habit*. Habits afford a nontrivial

[12] This is what Lucy (1996) calls "semiotic relativity". Jackendoff (1996), Clark (1998) or Carruthers (2002) argue for a non-trivial impact of language on thought, but avoid completely the question of diversity (be it linguistic or cognitive), or even argue against it.

[13] When confronted to a correlation of language and thought phenomena, the advocate of CIL has to show that this correlation is at least in part the result of an impact in the direction language → thought. He doesn't need to (although he can) argue for something as strong as the thesis that some forms of thought originate with language. Perhaps language only augments or facilitates some pre-linguistic capacities.

[14] Counterfactual reasoning is one of the classic areas of research, although the usual focus is on grammatical features such as if-clauses or verb tense/aspect/mood (*cf.* Bloom 1981), not on names.

but non-deterministic impact of language on thought. In the present case, the constant use of linguistic resources for counterfactual discourse (by hypothesis, present in every human language) arguably fosters counterfactual thought habits that show up even when we are not thinking for speaking (and, of course, also when we are thinking for speaking or for understanding other people's speech). A key point is that this questions the idea defended by Fodor or Searle that in every domain *original aboutness* belongs to mental representations while linguistic expressions have only *derived aboutness*. The idea is that only by internalizing social linguistic practices in which names intervene, humans acquire the cognitive tools that are involved in at least some human forms of thought.[15]

6. A diversity of linguistic diversities

Premise 2 is a necessary step in the argument from the claim that language affects thought to the conclusion that this impact is not homogeneous. The following series of quotes illustrates the possible stances on the topic of linguistic diversity in linguistics. I add a first stance that may be tempting for philosophers (we find it in the *Tractatus*). Chomsky's universalism has at least the restriction of human nature:

Linguistic hyper-universality
There are some features that we can expect to find in any language, natural or artificial, human, divine or alien.

Linguistic (Chomskyan) universality
"We can be pretty confident that the different stages that are attained by the language faculty are only different in a superficial fashion and that each one is largely determined by the common language faculty. The reason for believing that is pretty straightforward. It is simply that relevant experience is far too limited." (Chomsky 2000, 6)

---

[15] The role of speech in the socialization of thought constitutes Vygotsky's fundamental idea (Vygotsky 1962). The internalized language that is used as a cognitive tool is for him the public language and retains many of its public characteristics. That would affect the use of names when we think in inner speech.

Cross-linguistic diversity
"In actuality, no person speaks 'language in general' but always a particular language with its own characteristic structure of meaning." (Lucy 1996, 41)

Radical cross-linguistic diversity
"The more we discover about languages, the more diversity we find." (Evans & Levinson 2009, 436)

Intra-linguistic diversity (diversity of linguistic variants)
"Strictly speaking, nobody speaks a language; we all speak a linguistic variety." (Moreno Cabrera 2000, 47)

Idiolectal diversity, and Intra-individual linguistic diversity
"There aren't two people that speak exactly the same way. Even the same person doesn't speak the same way during her life, or in different moments of the same day." (Bernárdez 1999, 26)

What is the case with proper names and other types of expressions for which it has been argued that Kripke's view is correct? We've seen that Kripke rejects A), but most philosophers accept B), although for functional more than biological reasons. The more obvious form of LD would be cross-linguistic: as Pirahã and Malagasy perhaps show, different languages can incorporate different mechanisms of reference for some terms. But possibly there is also intra-linguistic diversity (of dialects, sociolects or idiolects), and even intra-individual diversity. It could be the case that a speaker, even a typical one, associates two conventions with a name, one descriptivist, the other causal-historical, and that he uses them in a flexible way according to the context. Some experimental philosophers have defended recently that natural kind terms and names are ambiguous between a descriptive and a causal-historical reading (Nichols *et al.* 2016).

7. Diversity in referential mechanisms and relativity of mental aboutness

A recurrent topic in the work of experimental philosophers is that the variability in intuitions is too anarchic. I don't accept that speakers'

intuitions are the only available evidence for a theory of reference. But the evidence of whatever kind could convince us that there is a great linguistic diversity in the referential mechanisms. Let's assume that this diversity originates in cultural life forms, but reaches the uses of individual speakers largely through their acquisition of a public language. We have then different ways in which the general cognitive impact of language could vary according to the distinct forms of LD. The most obvious form would be the one that arises from cross-linguistic diversity. Do the Pirahã have only a problem with the name "Jesus" or are they incapable of thinking about Jesus? I find it plausible that they have difficulties with Jesus' thoughts mainly because of the way they talk. Another possibility (that could account for the percentages detected by Machery *et al.*) is that in some languages there are two internal varieties, one causal-historical and the other descriptivist, or hybrid, or whatever. Then, one group could talk and think in one way and another group could talk and think in a different way. A third possibility is that there are members of the community that aren't well acculturated or have atypical intentions, that is, cases of idiolectal diversity or just incorrect use; this could be the cause of cognitive idiosyncrasies. Finally, if some terms are ambiguous, this could affect the corresponding concepts. As I said, some experimental philosophers have recently argued that natural kind terms and names are ambiguous (it is not clear if in English or in any language) with a descriptivist reading and a causal-historical one:

> […] our proposal is that natural kind terms (and plausibly names as well) are ambiguous, such that in some cases the reference is determined descriptively and in other cases the reference is determined non-descriptively (Nichols *et al*. 2016, 160)

If this were so, we could have here a case of cognitive impact similar to the one that affects the expression "time" and the concept TIME. We can choose to describe or to think about a situation as "a too *long* talk" or as "*wasting* too much time", depending on our use of the metaphor TIME IS SPACE or the other metaphor TIME IS MONEY, both present in the repertory of conventional metaphors that we acquire through the learning of English, according to our communicative or cognitive purposes. The second metaphor at least is absent in many cultures and languages (where

money doesn't exist), so we also have cross-linguistic diversity here. Analogously, Nichols *et al.* argue that depending on the context we can use a natural kind term causal-historically or descriptively. A defender of Premise 1 would only have to add the hypothesis that this makes us capable of a correlative cognitive flexibility in the mental use of our natural kind concepts (and of switching our metaphysics from realism about natural kinds to a more Lockean metaphysics).

With this move, experimental philosophers finish the exploration of the varieties of diversity that linguists have distinguished. In footnote 25 they are explicit with respect to a change from an emphasis on C) and E) to an emphasis on G):

> In an earlier paper […] two of us reject the assumption that there is a single set of reference intuitions in the population. In that paper, we allowed that different people might have intuitions that support different theories of reference […]. But we did not explore the possibility that within each of us, there are (at least) two ways of thinking about the reference of kind terms. (Nichols *et al*. 2016, 161)

Blanco Salgueiro (2017) points out that this kind of LD (intra-individual LD) can be used to avoid the radical view known as "linguistic determinism". It is possible for a language to contain versatile enough tools that respond to the current context in flexible ways. Be it plausible or not, Nichols *et al.*'s proposal suggests a new relativistic hypothesis: languages with two referential conventions associated with proper names and other types of expressions not only allow for using them in two different ways when speaking, but also using their mental correlates in different ways when thinking. These different conventions surely come ultimately from different cultural practices.

8. Conclusion

I have tried to show that there is an important connection between the philosophy of reference and the language-thought relation debate, in particular with the controversy about the plausibility of the linguistic relativity hypothesis.

The implications of the discussion are far-reaching. Our ability to share thoughts with people who speak other languages (and the possibility of translating our language into theirs) is not in question if the main impact of language on thought were universal. If the mechanisms for reference are shared cross-linguistically, then it can be argued for a semiotic impact of language on thought: perhaps the ability to think counterfactually about objects and kinds depends in part on language, but it happens that *any* language has resources that promote this ability. Arguably, nonverbal beings have little or no capacity for displaced thinking (not attached to the actual situation), and language can contribute to explain this human cognitive singularity.

The hypothetical impact of intra-individual diversity evidences the flexibility of human thought, its capacity to change its frames to respond to the actual context. If your ability to think of an individual descriptively or causal-historically depends on your having two linguistic conventions (and this may not be a language universal), this also makes you capable of understanding both uses of the terms, although there is a risk of misunderstanding with others (or even with yourself in different moments) if you are applying a different convention from your interlocutor; in a Gödel-type scenario the referent will be different, depending on which convention is applied.

In the case of cross-linguistic or cross-variant diversity, there can be systematic differences in the habitual ways of thinking about individuals and kinds due to linguistic diversity. This does not necessarily mean that the misunderstanding is insurmountable. If language influence is a question of promoting particular habits of thinking (as argued by current linguistic relativists), then you can grasp other ways of talking and thinking paying more attention, dedicating more mental resources, or using your imagination. And, of course, you can learn other languages, or new linguistic rules. Cognitive habits are reversible, but can also be persistent, so that you have to make an effort to understand and pursue ways of thinking you are not used to. Nevertheless, as the Jesus-case shows, some features in the language games can be so entrenched that it is near impossible for a speaker of a language to think of an individual or kind in a way not permitted by her language. For instance, because there is no place in Pirahã for Jesus' name, it is very difficult to find a place for Jesus in a pirahã's mind. Here, the question is not (as in the Gödel case) which should be the referent of

the name in a counterfactual scenario, but the very possibility to refer to an individual that existed far in the past.

All these possibilities are consistent with current research on the language-thought relation and the linguistic relativity hypothesis. This research has paid little attention to names and their potential cognitive impact. More empirical work on this topic (in particular, field work on the conventions for the use of names in many languages) is needed to properly answer the questions addressed.

What about Machery *et al.*'s position? I see it as a hypothesis on the influence of particular languages (and only indirectly of particular cultural practices) on a certain cognitive domain: the intuitions of ordinary speakers in imaginary scenarios such as the Gödel-case. This hypothesis is controversial for reasons alleged by their critics; there could be other factors that explain the difference in intuitions: the extraordinariness of the imaginary cases, the influence of folk theories or myths about language, etc. The focus on intuitions seems wrong, if we try to set the influence of the linguistic mechanisms for reference on the cognitive mechanisms for aboutness. We should focus instead on the differences in the ordinary use of referential terms to settle the question of linguistic diversity, and on the possible influence of these differences in cognitive tasks that involve mental reference (like counterfactual reasoning). But, of course, it is possible (in fact, I take it as a good hypothesis) that some of the differences in intuitions are due, in part, to differences in the conventions for the use of proper names.

Acknowledgments

I am grateful to two anonymous reviewers for their helpful remarks on a previous version of this paper. The work has been supported by the Spanish Ministry of Economy and Competitiveness in the framework of the research project FFI2014-52244-P.

References

BERNÁRDEZ, E. (1999): *Qué son las lenguas.* Madrid: Alianza.

BLANCO SALGUEIRO, A. (2017): *La relatividad lingüística (variaciones filosóficas).* Madrid: Akal.

BLOOM, A. (1984): *The Linguistic Shaping of Thought: A Study in the Impact of Language on Thinking in China and the West*. Hillsdale, New Jersey: L. Erlbaum.

CARRUTHERS, P. (2002): The Cognitive Functions of Language. *Behavioral and Brain Sciences* 25, 657-674.

CHOMSKY, N. (2000): *The Architecture of Language*. Oxford: Oxford University Press.

CLARK, A. (1998): Magic Words: How Language Augments Human Computation. In: Carruthers, P. & Boucher, J. (eds.): *Language and Thought: Interdisciplinary Themes*. Cambridge: Cambridge University Press.

CRYSTAL, D. (2000): *Language Death*. Cambridge: Cambridge University Press.

DEVITT, M. (2011): Experimental Semantics. *Philosophy and Phenomenological Research* 82(2), 418-435.

DEVITT, M. (2012): Whither Experimental Semantics? *Theoria, Revista de Teoría, Historia y Fundamentos de la Ciencia* 27(1), 5-36.

DEUTSCH, M. (2009): Experimental Philosophy and the Theory of Reference. *Mind & Language* 24(4), 445-466.

ENFIELD, N. J. (2015): Linguistic Relativity from Reference to Agency. *Annual Review of Anthropology* 44, 207-224.

EVANS, N. & LEVINSON, S. C. (2009): The Myth of Linguistic Universals: Language Diversity and Its Importance for Cognitive Science. *Behavioral and Brain Sciences* 32, 429-492.

EVERETT, D. (2005): Cultural Constraints on Grammar and Cognition in Pirahã. Another Look at the Design Features of Human Language. *Current Anthropology* 46(4), 621-646.

EVERETT, D. (2008): *Don't Sleep, there are snakes. Life and Language in the Amazonian Jungle*. London: Profile Books.

EVERETT, D. (2013): *Language: The Cultural Tool*. London: Profile Books.

GUMPERZ, J. J. & LEVINSON, S. C. (1996): *Rethinking Linguistic Relativity*. Cambridge: Cambridge University Press.

HANNON, M. J. (2017): Intuitions, Reflective Judgements, and Experimental Philosophy. *Synthese* http://doi.org/10.1007/s11229-017-1412-1.

HECK, J. G. (2017): Speaker's Reference, Semantic Reference, and Intuition. *Review of Philosophy and Psychology*, http://doi.org/10.1007(s13164-017-0362-3.

ICHIKAWA, J. MAITRA, I. & WEATHERSON, B. (2011): In Defense of a Kripkean Dogma. *Philosophy and Phenomenological Research* 85(1), 56-68.

JYLKKÄ, J., RAILO, H. & HAUKIOJA, J. (2009): Psychological Essentialism and Semantic Externalism: Evidence for Externalism in Lay Speakers' Language Use. *Philosophical Psychology* 22(1), 37-60.

JACKENDOFF, R. (1996): How Language Helps Us Think. *Pragmatics and Cognition* 4(1), 1-34.

KNOBE, J., BUCKWALTER, W., NICHOLS, S., ROBBINS, P., SARKISSIAN, H. & SOOMERS, T. (2012): Experimental Philosophy. *Annual Review of Psychology* 63, 81-99.

KRIPKE, S. (1972/1980): *Naming and Necessity*. Oxford: Blackwell.

KRIPKE, S. (1977): Speaker's Reference and Semantic Reference. *Midwest Studies in Philosophy* 2, 255-276.

LAM, B. (2010): Are Cantonese-Speakers Really Descriptivist? Revisiting Cross-Cultural Semantics. *Cognition* 115, 320-329.

LEVINSON, S. C. (2005): Comment to Everett (2005). *Current Anthropology* 46(4), 637-638.

LIAO, S. M. (2008): A Defense of Intuitions. *Philosophical Studies* 140, 247-262.

LUCY, J. A. (1996): The Scope of Linguistic Relativity: An Analysis and Review of Empirical Research. In: Gumperz & Levinson (eds.), 37-69.

LUDWIG, K. (2010): Intuitions and Relativity. *Philosophical Psychology* 23(4), 427-445.

MACHERY, E. (2012): Expertise and Intuitions about Reference. *Theoria, Revista de Teoría, Historia y Fundamentos de la Ciencia* 27(1), 37-54.

MACHERY, E., DEUTSCH, M., MALLON, R., NICHOLS, S., SYTSMA, J. & STICH, S. P. (2010): Semantic Intuitions: Reply to Lam. *Cognition* 117, 361-366.

MACHERY, E., MALLON, R., NICHOLS, S. & STICH, S. C. (2004): Semantics, Cross-Cultural Style. *Cognition* 92, B1-B12.

MACHERY, E., MALLON, R., NICHOLS, S. & STICH, S. C. (2013): If Folk Intuitions Vary, Then What? *Philosophy and Phenomenological Research*, 56(3) 618-635.

MACHERY, E., OLIVOLA, C. & DE BLANC, M. (2009): Linguistic and Metalinguistic Intuitions in the Philosophy of Language. *Analysis* 69, 689-694.

MACHERY, E., STICH, S., ROSE, D., CHATTERJEE, A., KARASAWA, K., STRUCHINER, N., SIRKER, S., USUI, N. & HASHIMOTO, T. (2015): Gettier Across Cultures. *Noûs* 51(3), 645-664.

MACHERY, E., SYTSMA, J. & DEUTSCH, M. (2015): Speaker's Reference and Cross-Cultural Semantics. In: Bianchi, A. (ed.): *On Reference*. Oxford: Oxford University Press, 62-76.

MALLON, R., MACHERY, E., NICHOLS, S. & STICH, S. (2009): Against Arguments from Reference. *Philosophy and Phenomenological Research* 59(2), 332-356.

MARTÍ, G. (2009): Against Semantic Multi-Culturalism. *Analisis* 69(1), 42-48.

MORENO CABRERA, J. C. (2000): *La dignidad e igualdad de las lenguas. Crítica de la discriminación lingüística*. Madrid: Alianza.

NADO, J. & JOHNSON, M. (2016): Intuitions and the Theory of Reference. In: Nado, J. (ed): *Advances in Experimental Philosophy and Philosophical Methodology*. New York: Bloomsbury, 125-154.

NICHOLS, S., PINILLOS, N. A. & MALLON, R. (2016): Ambiguous Reference. *Mind* 125(417), 145-175.

NISBETT, R. E. (2003): *The Geography of Thought: How Asians and Westerners Think Differently … and Why*. New York: Free Press.

NISBETT, R. E., PENG, K., CHOI, I. & NORENZAYAN, A. (2001): Culture and Systems of Thought: Holistic vs. Analytic Cognition. *Psychological Review* 108, 291-310.

OCHS, E. (1976): The Universality of Conversational Postulates. *Language in Society* 5(1), 67-80.

PUTNAM, H. (1975): The Meaning of 'Meaning'. In: Gunderson, K. (ed.): *Language, Mind and Knowledge, Minnesota Studies in the Philosophy of Science VII*. Minneapolis: University of Minnesota Press, 131-193.

STICH, S. & TOBIA, K. (2016): Experimental Philosophy and the Philosophical Tradition. In: Buckwalter, W. & Sytsma, J. (eds): *Blackwell Companion to Experimental Philosophy*. Oxford: Blackwell.

STRAWSON, P. (1959): *Individuals*. London: Methuen.

SYTSMA, J., LIVENGOOD, J., SATO, R. & OGUCHI, M. (2015): Reference in the Land of the Rising Sun: A Cross-Cultural Study on the Reference of Proper Names. *Review of Philosophy and Psychology* 6, 213-230, doi: 10.1007/s13164-014-0206-3.

VAESEN, K., PETERSON, M. & VAN BEZOOIJEN, B. (2013): The Reliability of Armchair Intuitions. *Metaphilosophy* 44(5), 559-578.

VYGOTSKY, L. (1962): *Thought and Language*. Cambridge, Mass.: MIT Press.

WEINBERG, J., NICHOLS, S. & STICH, S. (2001): Normativity and Epistemic Intuitions. *Philosophical Topics* 29(1&2), 429-460.

WITTGENSTEIN, L. (1953/1958): *Philosophische Untersuchungen / Philosophical Investigations*. Oxford: Blackwell.

WHORF, B. L. (1939): The Relation of Habitual Thought and Behavior to Language. In: Whorf, B. L. (1972): *Language, Thought and Reality* [1956]. Cambridge, Mass.: MIT Press, 134-159.

Organon F 25 (4) 2018: 564-588

Returning to a Tension within Grice's Original Account of Nonnatural Meaning

KONSTANTY KUZMA[1]

ABSTRACT: It has become a commonplace to regard Grice's project in "Meaning" as plagued by circularity, and almost as prevalent to dismiss such charges as unfounded. Much of the controversy surrounding Grice's presumed circularity revolves around the question whether Grice is committed to a reductionist project of meaning, or whether it is merely meant to elucidate the nature of meaning without pretending to reduce it to something meaningless. Rarely, however, are these views developed as part of a systematic analysis of Grice's original paper, as this paper seeks to do. My paper consists of two parts. In the first part, I try to show how Grice can be defended from John Searle's criticism relating to the famous American soldier example and argue that Searle's suggested amendments run counter to Grice's ambitions. In the second part of my paper, I illustrate – drawing on the first part – why "Meaning" both makes it necessary and seem impossible that the timeless meaning of utterances be fully reducible to individual utterances and thus to individual speakers' intentions. I argue that this seriously challenges the view that Grice is putting forward a theory of intention-based semantics in "Meaning" which would present a viable alternative to later developments of his theory.

KEYWORDS: Grice – intentions – meaning – pragmatics.

[1] Received: 6 February 2018 / Accepted: 15 September 2018

✉ Konstanty Kuzma

Department of Philosophy
Ludwig Maximilian University Munich
Geschwister-Scholl-Platz 1 (A 221), 80539 Munich, Germany

e-mail: k.kuzma@campus.lmu.de

0. Introduction

This paper is first and foremost intended to return to and bring out a tension within Grice's seminal "Meaning". While the tension has previously been observed (it is explicitly stated in Strawson 1971 and Burge 1979), it is rarely formulated within the context of a systematic treatment of Grice's original theory of meaning, and to my mind never against the background of its wide implications. There seem to be two principal reasons for this situation. One is that Grice and Schiffer soon developed an alternative way of pursuing a broadly Gricean approach that does not rely on the self-referential intention of "Meaning" and is widely regarded to be the more promising path towards constructing an intention-based semantics (see Grice 1989f; Schiffer 1972). The other is that Grice has often been disassociated from the attempt to fully reduce the semantic to the psychological. As Avramides has argued at length, one can conceive of Grice's project as one of mutual elucidation rather than one-way reduction (cf. Avramides 1989, ch. 1). With this possibility in mind, one can return to Grice's original account from "Meaning" without falling prey to the tension I am about to delve into.

My paper does not pursue either of these approaches. In fact, it is orthogonal to the adequacy of the Schiffer/late Grice approach, and only relevant to the anti-reductionist insofar as it (implicitly) disassociates her from the project of an intention-based semantics. This is because the paper puts pressure on the third alternative the above landscape leaves open, which is to return to the original account of "Meaning" with a reductionist project in mind.[2] I will argue that this is both the most natural way to read that original paper (Section 1), and the only hope of reconstructing "Meaning" as an account of intention-based semantics (Section 2). Because of this limited perspective on Grice's larger project, I will mainly draw on John Searle, who has argued that the self-referential intention from "Meaning" is key to a proper account of meaning (cf. Searle 2007, 14).[3] My aim will be to show that Grice's original approach is inconsistent as an attempt

---

[2] Thus, I pursue what Avramides refers to as a "strong, reductive interpretation" (Avramides 1989, 13).

[3] This is so even though Searle also thinks Grice confuses the explanatory role self-referentiality should play in such an account.

to construct an intention-based semantics, so that pace Searle, we should not regard the self-reflexive intention as key to the Gricean project.[4]

The tension my paper is concerned with is the following. As I will argue, "Meaning" both makes it necessary and seem impossible that the timeless meaning of utterances be fully reducible to individual utterances and thus of individual speaker's intentions.[5] This is because for Grice to provide a theory of meaning – specifically, an intention-based semantics –, the explicatory dependency between timeless meaning and speaker's meaning must be one-way (this is the requirement meant to be brought out by the mildly dramatic talk of "full" reducibility as opposed to reducibility simpliciter above).[6] That is, for Grice's project to succeed as an instance of intention-based semantics, timeless meaning must be analyzable in terms of speaker's meaning without semantic remainder.[7] The fact that "That book is green" means what it means, for instance, must solely be a function of a community of speakers intending it to mean what it means. Once this requirement is brought into view, another difficulty arises, which is that there seems to be no way of meaning anything complex by one's utterances independent of the existence of timeless meaning. For example, there is no hope of meaning that the book over there is green without there being a set of conventions which fix the

---

4 The self-reflexive intention of "Meaning" has been confronted with concerns about its presumed circularity and implausibility, prompting the development of alternative approaches to the intended effect of an utterance. Cf. Neale (1992, 548); Recanati (1986); Sperber & Wilson (1986). I will argue that even if we grant Grice that there is nothing circular or implausible about the self-reflexive intention, one cannot both hold on to Grice's original account from "Meaning" and pursue the project of an intention-based semantics. It is in this sense that the paper is meant to discredit Grice's original account as a viable intention-based alternative to later versions of the theory.

5 Strawson acknowledges this tension but thinks that you need not posit full reducibility from a Gricean perspective (cf. Strawson 1971, 174). I discuss Strawson's solution in Section 2.

6 Drawing on Grice's characterization of both terms, I use "speaker's meaning" to denote the meaning intended by the speaker in uttering an utterance, whereas "timeless meaning" denotes the conventional meaning of an utterance.

7 I borrow this way of framing the requirements of Grice's theory from Grandy & Warner (2017). This sets my reading of Grice's original paper apart from interpretations that take Grice to merely aspire a reductionist project in the sense of conceptual elaboration (see e.g. Neale 1992).

meaning of a set of signs roughly synonymous to "That book is green". And that would of course mean that whatever the account presented in "Meaning" amounted to, it would not serve the reductionist project that Grice is aiming at. For that, again, would require timeless meaning to be fully reducible to speaker's meaning without semantic remainder.[8]

I will approach the said tension through a discussion of two lines of criticisms that Searle has raised vis-à-vis Grice's conception of meaning. Though I share Searle's verdict that Grice fails to provide a theory of meaning, I share it for different reasons, meaning that the discussion of Searle will lead up to my criticism of Grice in a roundabout fashion. I will first rehearse Searle's criticism of Grice, then try and show how Grice can be defended against it, and finally argue that Searle's objections and his American soldier example (to which I will get shortly) can nevertheless help us see what is fundamentally problematic with the conception that Grice offers in "Meaning".

1. Searle's criticism

I will begin my discussion of John Searle's criticism of Grice with the arguments put forward in *Speech Acts*. Though Searle later revised the position argued for in *Speech Acts*, it will be helpful to briefly return to it. The object of Searle's inquiry in *Speech Acts*, then, is the central definition of nonnatural meaning argued for in "Meaning", which Searle cites in the following fashion:

> To say that a speaker *S* meant something by *X* is to say that *S* intended the utterance of *X* to produce some effect in a hearer *H* by means of the recognition of this intention. (Searle 1969, 43)

---

[8] I stress Grandy and Warner's "without semantic remainder" proviso (see Grandy & Warner 2017) because there is an obvious sense in which Grice reduces all meaning to intentions. After all, speaker's meaning is constituted by intentions, while timeless meaning is nothing but regularities among those very intentions. The issue I will discuss towards the end of section 2 is that while intentions determine the individual meaning of utterances, they can only do so against the context of an already existing set of conventions.

It will be useful to have a shorthand for this definition, so let us call it Meaning^{NN}. The purpose of Meaning^{NN}'s self-reflexivity ("by means of the recognition of this intention") is to separate cases of nonnaturally meaning something from cases in which one intentionally produces an effect in someone without one's intention playing any part in the production of that effect. The latter case, Grice argues, would not be a case of nonnaturally meaning anything. An example for this case cited in "Meaning" is the scene of Herod presenting Salome with the head of John the Baptist (Grice 1989b, 218). While Herod intended to make Salome believe that John the Baptist has died by producing the latter's head on a platter, it is not the case (or so argues Grice) that Herod *meant* anything by showing her the head of Salome. Herod's intention to make Salome believe that John the Baptist is dead does not play a role in producing the effect of her believing that John the Baptist is dead. If, on the other hand, Herod had (to the unquestionable detriment of art history) relied on less dramatic means of getting the message across and simply *said*, "I had John the Baptist killed", his intention of getting her to know that John the Baptist is dead would have indeed played a part in producing that effect. Therefore, the latter case would have been a case of nonnatural meaning.[9]

Searle takes issue with this account for two reasons. One is that Grice does not account for the way that meaning "can be a matter of rules or conventions" (Searle 1969, 43). In other words, Searle claims that "Meaning" does not acknowledge the way in which meaning something by one's utterance is connected to what that utterance usually means. The other is that Grice is supposed by Searle to be wrong about the intended effect of utterances. While "Meaning" states that nonnaturally meaning something by one's utterance (in the case of indicative sentences) is an instance of intending to "induce by x a belief in an audience" (Grice 1989b, 219), Searle thinks that it is merely an instance of producing understanding on the hearer's part. Since the latter objection is developed in Searle's recent paper "Grice on Meaning: 50 Years Later", the discussion of which I will

[9] It has been debated whether Grice's intuitions are correct concerning the contrast between natural and nonnatural meaning in the Herod example. In particular, it is controversial whether the self-reflexive clause (which posits that the intention to produce an effect must itself be intended to function as a reason for producing that effect) is needed. Cf. Neale (1992, 548); Recanati (1986); Sperber & Wilson (1986).

take up shortly, I will – for the time being – concentrate on the first objection, i.e. that Grice's account fails to account for the connection between speaker's meaning and timeless meaning, which brings us to Searle's famous American soldier example.[10]

The example goes as follows. We are supposed to imagine that an American soldier who has been captured by Italian troops is trying to make his captors believe that he is a German officer. Knowing virtually no Italian or German, he puts on a show to tell them that he's a German officer by reciting the only German line that he knows: "*Kennst du das Land, wo die Zitronen blühen?*" Searle maintains that the soldier's utterance does not mean either "I am a German officer" or that utterance's German-language equivalent "*Ich bin ein deutscher Offizier*". But the Gricean analysis, he thinks, not only implies that this is what it means, but that furthermore it follows that "any sentence can be uttered with any meaning whatever, given that the circumstances make possible the appropriate intentions" (Searle 1969, 45). To prevent meaning from being fixed arbitrarily, Searle suggests incorporating the conventional meaning of utterances into Grice's account of meaning. Thus, Searle arrives at the following, amended version of Grice's account of meaning:

> In our analysis of illocutionary acts, we must capture both the intentional and the conventional aspects and especially the relationship between them. In the performance of an illocutionary act in the literal utterance of a sentence, the speaker intends to produce a certain effect by means of getting the hearer to recognize his intention to produce that effect; and furthermore, if he is using words literally, he intends this recognition to be achieved in virtue of the fact that the rules for using the expressions he utters associate the expression with the production of that effect. It is this *combination* of elements which we shall need to express in our analysis of the illocutionary act. (Searle 1969, 45)

---

[10] The example is presented in Searle (1969, 44f). Notable (and for the most part dismissive) discussions of the example are to be found in Grice (1989f); Armstrong (1971, 440-441); Bennett (1973, 164-165); Martinich (1984, 122-125); Schiffer (1972, ch. 2); Yu (1979, sct. 3).

This establishes the connection between speaker's meaning and timeless meaning that Searle's above-mentioned criticism of Meaning^{NN} called for. Meaning something by one's utterance is not a completely arbitrary bestowal of meaning on an utterance that can by that act be made to mean anything. Rather, Searle thinks, "what we can mean is at least sometimes a function of what we are saying" (Searle 1969, 45). Citing Wittgenstein, Searle reminds us that you cannot say "it's cold here" and *mean* the opposite (Searle 1969, 45).

There are several ways of responding to Searle's criticism, some of which Grice himself pointed to in "Utterer's Meaning and Intentions". Before I discuss some of those responses, however, it is worth pausing for one moment to deliberate Searle's counter-suggestion which he thinks provides a way of both avoiding counter-examples of the sort exemplified by the American soldier and establishing a connection between speaker's meaning and timeless meaning. Even if we set aside the problem posed by counter-examples for a moment, it is quite clear that Searle's suggestion for amending Grice's account of meaning will not do as far as Grice's project is concerned. This is because the connection that Searle establishes between an utterance and its conventional meaning makes it impossible to arrive at a reductive account of meaning. Searle suggests that literal utterances be thought of as resulting from a combination of the utterer's intention and his utterance's conventional meaning, so that the analysis of meaning includes the very thing that is supposed to be explained. In other words, we have arrived at an explanation of meaning which itself includes a reference to meaning in the form of "rules for using the expressions he [the speaker] uses". Initially, it is not entirely clear *what* kind of nonnatural meaning Searle is attempting to give an account of – whether it is of an utterance's timeless meaning or of speaker's meaning. But in the former case, the account would be blatantly circular, as Searle would be analyzing an utterance's timeless meaning in terms of the utterer's intentions and the utterance's timeless meaning. Even if we are more charitable towards Searle and allow that he is attempting to provide a definition for speaker's meaning, whereas the "rules for using the expressions" are clearly a reference to timeless meaning, the problem remains standing that in trying to account for meaning, he is making recourse to something that is already meaningful (namely the rules for expression use). While not strictly speaking circular, Searle would still not be providing a proper account of what makes

utterances meaningful. If we have conventional meaning to fall back on in uttering literal utterances, we do not need intentions to make them meaningful – they already are.[11]

It will be useful to keep this in mind as Searle's misconception of Grice's aim in "Meaning" shapes his entire discussion of it. I now turn back to the American soldier example to show how Grice can deal with it. To recap, the example was supposed to pose a problem for Grice because his analysis would suggest that "*Kennst du das Land, wo die Zitronen blühen?*" could be brought to mean "I am a German officer", whereas of course it means something completely different. Now as Grice rightly points out, it is quite difficult to imagine how the American soldier could want to get his captors to think that the *words* he is uttering mean "I am a German officer" (Grice 1989f, 101-102).[12] It would be much more natural to describe the example analogously to the Herod case, so that the Italian troops merely infer, from the observed circumstances of the soldier's utterance (i.e. independently of his intentions) – his German-sounding words and the fact that he has the demeanor of a representative of the military – that he is a German officer.[13] And if this were indeed the proper way to describe the scenario, then (again analogously to the Herod example) we would not be dealing with a case of nonnatural meaning. For even if the American soldier had intended his captors to go through the said inferential steps, his intention could not have been supposed to play a role in their arriving at the belief that he is a German officer.

The argument could have ended here. As Grice himself observes, this is the most intuitive way of describing Searle's example, and Grice's response seems both ample and satisfactory. But charitable as he is, Grice

---

[11] It is worth noting that intentions appear to do no work on the latter reading of Searle's account of the meaning of literal utterances. For what does it matter what *I* mean by a literal utterance if its meaning is already fixed by the conventional meaning? Martinich argues that because of such constraints, Searle effectively ties utterances to their conventional meaning (Martinich 1984, 124).

[12] The same issue is raised in Armstrong (1971, 440); Bennett (1973, 164). Schiffer goes even further in questioning whether the American soldier meant anything at all by his utterance. See Schiffer (1972, 27).

[13] This is the first of two possible reinterpretations of Searle's example offered in Schiffer (1972, 28). Also see Grice (1989f, 101).

allows that Searle have his way. He assumes with Searle that the American soldier in fact wants the Italian troops to come to believe that he is a German officer "via a belief that the words which he uttered were the German for 'I am a German officer'" (Grice 1989f, 101). And if this is something we can imagine, Grice continues, then we should say that the American soldier meant by "*Kennst du das Land, wo die Zitronen blühen?*" that he is a German officer. Does this sound counter-intuitive? Hardly so, for Grice explicitly denies the implication that this is what the German officer means by the *words* "*Kennst du das Land, wo die Zitronen blühen?*" (Grice 1989f, 102). The relevant analysis presented in "Meaning" is merely intended to bring out what a Speaker S means in *uttering* a sentence X. There is no good inference to the commitment on Grice's part that that is what the sentence normally means. And of course, we can mean something in making an utterance that departs from its conventional meaning. One need not refer to Grice's theory of conversational implicature to acknowledge this point. Even someone who objected to that theory's logical ramifications must surely acknowledge that departure from the normal meaning of one's utterances is something we do on a daily basis.

Imagine the following scenario: a group of friends meet in a bar to have a drink. When the waiter comes by to take everyone's order, one of the friends misspeaks while ordering a beer, saying "bear" instead of "beer". The group breaks out in laughter, and the waiter picks up the slip of the tongue, reacting with a dry joke which produces further laughter among the group. The next time that the waiter comes by to take orders, the friends order "bears" rather than "beers". It seems quite natural to describe the friends as intending to order "beers" when asking to get "bears" from that moment on, and to expect the waiter to understand their cue. Still, there is no good inference to saying that this is what the word "bear" means. In fact, it is precisely due to it not being the word's conventional meaning that it provokes laughter among the group.

The mistake on Searle's part is to assume that Meaning^{NN} is supposed to do all the work for Grice's theory of meaning. When Grice is saying that in uttering a sentence X, S intends to produce an effect in hearer H in virtue of H's recognition of that intention, he is giving an account of speaker's meaning. In other words, he is trying to give an account of what happens when someone means something by an utterance. But this is not to say that a speaker's meaning something by an utterance fully accounts for the

utterance's meaning. In particular, it does not account for the utterance's timeless meaning. Speaker's meaning only accounts for what a speaker means by uttering an utterance.

In certain cases, this may be all that matters. In the American soldier example (assuming that it is an instance of nonnatural meaning), it is quite irrelevant that the sentence uttered by the soldier really means "Knowest thou the land where the lemon trees bloom?", as its conventional meaning stands in no relation to the meaning the speaker intended to convey in uttering that sentence in German. But in most cases of verbal communication, the command of the spoken language will be essential to deciphering the intended message. As Grice illustrates in "Logic and Conversation", this is even true of cases of nonliteral speech, as the ability to understand its meaning requires an understanding of the literal meaning of an utterance just as much as instances of literal speech do (Cf. Grice, 1989a, 30f). It is only through an understanding of an utterance's literal meaning that one arrives at the intended meaning of a non-literal utterance, for one must first pass through it and its inadequacy when interpreted literally (together with an application of the Cooperative Principle and possibly the Maxims of Conversation) to be able, according to Grice, to arrive at a proper understanding of how to take the message instead.[14]

This is also why, pace Searle, one cannot arbitrarily fix the meaning of utterances. Even when an utterance is meant non-literally, the literal meaning of the utterance plays a part in arriving at its meaning.[15] Grice is very explicit on the so-called M-intentions (that is meaning intentions) being bound by what can be transferred in an act of communication (this is part of the reason why he repeatedly refers to conversations as a rational

[14] As has been pointed out to me, the literal adequacy of certain metaphorical utterances (e.g. "No man is an island") puts pressure on the view that one should regard the inadequacy of an utterance when interpreted literally as a necessary point of departure for interpreting non-literal utterances. Arguably, "inadequacy" may thus be too strong a word, though the question of the exact formulation of the process of getting from literal to non-literal meaning is not entirely relevant to the purpose of this paper. The important point is that communication is rationally constrained (in part by timeless meaning).

[15] This is the principal reason why Neale objects to the claim that Grice neglects the role that timeless meaning plays in working out communicative intentions. Cf. Neale (1992, sct. 6).

endeavor) (cf. Grice, 1989a, 31). It is not the case that one can utter, in the middle of an unrelated conversation, "blob" with the intention of producing in one's hearer the belief that "Lewis Carroll is the pen name of Charles Lutwidge Dodgson". Rather, one can mean by one's utterances what they normally mean as well as what can be conveyed by them non-literally by way of drawing upon the principles of conversation laid out in "Logic and Conversation". And this bars one from being able to mean anything by anything whatsoever.[16]

There is of course a trivial sense in which it is indeed the case that one can mean anything by anything whatsoever, namely that we can imagine circumstances under which any sentence could be used to mean just about anything. Surely, we do not want to deny that under specific circumstances, say during a game, we could stipulate or otherwise imply that utterances mean something else than what they normally mean and be almost unlimited in our freedom to do so. But this does not imply that you could mean anything by uttering any sentence solely in virtue of your intending it to mean something, as in the Lewis Carrol example above, nor does it imply that this is what the words the utterance consists of mean. It is true that Searle thinks it is a problem that you can in principle mean anything by anything even though he acknowledges that this is only true given "that the circumstances make possible the appropriate intentions" (Searle 1969, 45). But this is only because Searle a) does not acknowledge the way in which circumstances put a rational constraint on what you can mean by an utterance, and because b) he does not realize that Grice is speaking about an utterance's speaker's meaning, and not its timeless meaning (cf. Armstrong 1971, 440-441; Bennett 1973, 164-165). Once we appreciate the latter two reservations, the charge becomes harmless. Grice is not committed to the absurd view that you can, under any given circumstances, mean "Lewis Carroll is the pen name of Charles Lutwidge Dodgson" by uttering anything whatsoever. His view is rather that the obtaining of appropriate circumstances allows you to utter that sentence while meaning something other than what is normally meant by its utterance. Furthermore, by that act, the sentence will not suddenly change its timeless meaning, but will

[16] The extent of this rational constraint is dangerously downplayed in Martinich (1984, 122-125). See Neale (1992, especially scts. 5 & 6), for an exposition of Grice's need for and deliverance of rational constraint on what an utterer can mean.

instead help instantiate a different utterance's meaning than that which is normally associated with its utterance.

This addresses the first line of critique raised in *Speech Acts*. It is not the case that Grice makes no connection between utterances and their timeless meaning (or conventional meaning), as one in general needs to know the literal meaning of utterances to be able to decipher even their non-literal meaning (as I pointed out above, the American soldier example – when interpreted as a case of nonnatural meaning – is a notable exception). This is why Grice repeatedly stresses that one is generally assumed to be intending to convey the literal meaning of one's utterances, which assumption is only dropped if it cannot be reconciled with the speaker's observing the Cooperative Principle (cf. Grice 1989b, 222; Grice 1989a, 30f). But even such deviance from the literal meaning of one's utterances and the arriving at their non-literal meaning is rule-governed, so that Searle is wrong in claiming that his American soldier example shows that one can, on a Gricean picture, mean anything by uttering any sentence whatsoever (lest we mean by that that we can imagine appropriate circumstances under which any given utterance could be brought to mean anything, which implication, for the reasons mentioned above, would not be problematic).

What about the second line of critique raised by Searle in *Speech Acts*, i.e. the charge that Grice is wrong about the intended effect of utterances? As I already mentioned, Searle altered and amended his critique from *Speech Acts* in his recent paper "Grice on Meaning: 50 Years Later", so that it will be useful to look at both texts to get an idea of Searle's position. To reiterate, in *Speech Acts* Searle objected to Grice's contention that nonnaturally meaning something by an utterance is an instance of trying to induce a belief in an audience. As Searle points out, we can utter a sentence and nonnaturally mean something by it without having any intention of inducing a belief in our audience (Searle 1969, 46-48). Even if we restrict our attention to indicative sentences, an analysis of which Grice was chiefly attempting to give in "Meaning", this charge seems justified. Take the following promenade example: if I take a stroll with a friend and remark on the beauty of the surroundings by saying "What a remarkable landscape this is!", it seems wrong to suggest that I am thereby attempting to get my audience to *believe* that the landscape is remarkable. Nor does it appear right to say, as Grice later claimed when responding to criticisms of this sort (Grice 1989e, 123), that I am trying to inform my friend of *my own*

belief about the remarkableness of the landscape.[17] Indeed, it seems wrong to describe my utterance as being primarily concerned with beliefs at all.[18]

Searle's suggestion in *Speech Acts* was that the proper way to describe someone's nonnaturally meaning something by an utterance is to say that the speaker is thereby intending to produce understanding on the hearer's part. Hence, in the promenade example I am merely getting my friend to understand what it is I am trying to say. In Searle's analysis, there is no reference to beliefs anymore, but only to getting my hearer to know what it is I am trying to communicate. But Searle later admitted that it was possible to nonnaturally mean something by one's utterance without even intending to produce understanding in one's audience (Searle 2007, 13f). A standard example is soliloquy, which also does not seem to be an example of producing understanding in one's hearer even though it is clear that one nevertheless nonnaturally means something by one's utterances.[19] The lesson Searle draws from this in "Grice on Meaning: 50 Years Later" is that Grice confused his account for an account of meaning, when in fact he was giving a (flawed) account of communication, and that Searle's analysis in *Speech Acts* followed Grice in this. Importantly, Searle still contends in "Grice on Meaning" that his *Speech Acts* account is superior to that of Grice because it can deal with all cases of nonnatural meaning save for

---

[17] Of course, Grice's attempt to save his analysis by suggesting that it is one's own beliefs that one is attempting to convey when nonnaturally meaning something by an utterance can best be challenged by finding examples in which one *is* in fact attempting to get the audience to believe something, as Grice's earlier analysis suggested was always the case. If, for instance, I am having a lengthy argument about youth unemployment in Europe with a friend, and after half an hour I verbally produce a statistic which I think will be devastating for my friend's position, it is clear that in producing the statistic I am not getting him to think what my position is.

[18] There is extensive literature on the problem of conceptualizing an utterance's intended effect. See Lycan (2008, 89-91); Neale (1992, sct. 5); Schiffer (1972, ch. 3); Strawson (1971, 172-173).

[19] See Ziff (1967) and Vlach (1981, 384-386), for useful expositions why audienceless cases constitute a problem for Grice (and, by extension, for Searle as well). Grice invokes the audience counterfactually to deal with the problem (so that one should understand the utterer as intending that were there an audience, the intended effect would be brought about), while Schiffer argues that at least in certain cases of soliloquy the speaker is his or her own audience. Cf. Grice (1989f, sct. 5); Schiffer (1972, 80).

those in which the speech act performed is defective (for lack of an intention to produce understanding), whereas there are perfectly nondefective speech acts (such as the promenade example) which Grice's analysis cannot deal with. The fact remains, however, that the accounts given in "Meaning" and *Speech Acts* are unable to deal with cases in which someone nonnaturally means something by one's utterance without intending to produce any effect in one's audience, and that according to Searle they provide accounts of communication rather than meaning as a result (Searle 2007, 14).

Given Searle's above-mentioned confusion about the aim of "Meaning", it is no surprise that his revised account of meaning in "Grice on Meaning" again fails to provide the kind of theory of meaning that Grice was looking for. Searle's new suggestion is to think of the literal meaning of indicative sentences in truth-functional terms:

> The meaning intention consists in the *intentional imposition of conditions of satisfaction* (in the sense of requirement) on conditions of satisfaction (in the sense of things required). The initial condition of satisfaction is simply that I produce the utterance, but the distinction between the utterance without meaning it, and the meaningful utterance where the meaningfulness is intended, is that the utterance itself, the condition of satisfaction of my intention to produce that utterance, has further conditions of satisfaction. In this case [i.e. a literal, indicative sentence] it has truth conditions. […] Analogous remarks can be made about directives and other forms of speech acts. Thus if I utter the French sentence "Fermez la porte" without meaning it, but just, for example, as practicing French pronunciation, the condition of satisfaction of my intention in action is simply that the intention in action should produce that utterance. But if I not only utter it but mean it, that is, mean it as a directive, then the conditions of satisfaction include that the hearer close the door. (Searle 2007, 15f)

We can leave aside any worries about whether one can really distinguish as easily between cases of saying and meaning something and saying something without meaning it. The important thing is that the impossibility of producing a reductive account of meaning reemerges. For Searle again explains meaning in terms of an act which is already meaningful, namely the intentional imposition of truth functions on an utterance. The articulation

and imposition of truth conditions on a series of sounds requires that those sounds be adequate vehicles for the transfer of those truth conditions. Otherwise the problem of being able to mean literally anything by any utterance whatsoever would emerge, as one could indeed be able to utter "blob" and mean "Lewis Carroll is the pen name of Charles Lutwidge Dodgson" just in virtue of one's wanting it to mean just that. In truth, we find that we are dependent on the timeless meaning of utterances for them to have truth conditions. "It is raining" is true if it is raining because that is what it means, and not just in virtue of my wanting it to be true under those conditions of satisfaction. In other words, Searle again ends up explaining meaning based on something that already presupposes meaning, namely the intentional imposition of truth functions on an utterance.

To be fair, in "Grice on Meaning" Searle does not pretend to be fixing Grice's account of meaning, but instead declares that his account is "Gricean in spirit" (Searle 2007, 17). But even his contention that Grice mistakes his account for a theory of meaning is to be handled with care. As I already hinted at, it is not the case that Meaning^{NN} is supposed to do all the work on Grice's account of meaning. That is why, granted that it may be better to describe the intention behind nonnaturally meaning something by an utterance as wanting to produce understanding rather than belief or belief communication, it is no real threat to Grice's project to state that Meaning^{NN} provides an account of "communication" as opposed to "meaning", so long as one means by this that Grice is giving an account of what it means for someone to nonnaturally mean something by an utterance, and not, as Searle wrongly suggests, an account of what utterances normally mean (more on this below).

2. Grice's account and a fundamental weakness

We said that Grice's account of meaning could be summarized in the following fashion: to say that a speaker S meant something by X is to say that S intended the utterance of X to produce – to incorporate Searle's suggestion – understanding in a hearer H by means of the recognition of this intention. This, again, is meant to account for what a speaker means in uttering an utterance. In other words, Grice is presenting an account of speaker's meaning.

As we saw, speaker's meaning does not necessarily correspond with what the uttered words normally mean. Indeed, you can nonnaturally mean something by uttering a series of sounds that has no conventional meaning in any language whatsoever (even though this requires that appropriate circumstances obtain). Normally however, you do need a grasp of the literal and conventional meaning of a sentence to unpack the utterance – whether it is literal, or not. If Searle were right and Grice were attempting to account for meaning solely in virtue of Meaning^{NN}, then Grice too would not be offering a reductive account of meaning. For even if we granted Grice that one cannot mean anything whatsoever by any odd utterance but must instead arrive at an understanding of speaker's meaning by way of the utterance's literal meaning, his account would fail as a theory of meaning for the same reasons as Searle's counter-suggestions did if he simply presupposed the existence of literal meaning.

But Grice has a separate story to tell about timeless meaning, which he thinks is indeed reductive. This is that the timeless meaning of an utterance is a function of what speakers in a linguistic community mean by that utterance, a view Brandom has called "regularism" (Brandom 1994, 26-30). According to regularism, rules guiding the use of an expression – on a rule-based account, its meaning – are nothing other than a description of the regularities pertaining to the use of those expressions within a linguistic community. In other words, to talk about norms "is just to talk about regularities" (Brandom 1994, 27), a view which Kripke famously attributes to Wittgenstein (cf. Kripke 1982, sct. 3). This yields a complex picture of Grice's project of an intention-based semantics. On the level of individual utterances, an utterance means what a speaker intends it to mean. But it is regularities among just this kind of M-intentions which yield the timeless meaning of utterances. And it is of course the latter which figure as a constraint on what speakers can intend by individual utterances.

Does this important aspect of Grice's theory save the project of an intention-based semantics as presented in "Meaning"? It would if it allowed us to reduce timeless meaning to speaker's meaning without semantic remainder. But as the discussion of Searle's American soldier example brought out, Grice's account of speaker's meaning cannot be conceptualized independently of timeless meaning. Recall that for a speaker to nonnaturally mean something by his utterance, and for a hearer to understand it, both usually rely on the literal meaning of the utterance (a notable

exception are utterances which are not based on or even related to actual words). Thus, if I make the utterance "It is raining", my hearer will normally have to know what the utterance usually means to be able to understand what I mean by uttering it.[20] This is evidently so in the case of literal utterances, because the timeless meaning of "It is raining" would then be precisely what I am trying to communicate. But it is also true if my utterance were meant non-literally, in which case the hearer would have to draw upon the Cooperative Principle (and possibly the Maxims of Conversation) to infer, according to the principles laid down in "Logic and Conversation" what I mean in uttering the sentence "It is raining". As I laid out above, this is the reason why speakers are generally assumed to be intending to mean their utterances literally. Now this reliance on the literal meaning of utterances brings up the same problem that was earlier put forward against Searle. For a speaker to be able to mean something by his utterance, he relies on the timeless meaning of it independently of the question whether it generally communicates what he is intending to convey. It does not matter that the timeless meaning of utterances derives from members of a linguistic community intending to mean something by it. For those speakers, in having meant something by their utterances, themselves relied on timeless meaning in uttering the utterances which contributed to the utterance's meaning what it means today. So that Grice, like Searle, is unable to explain meaning without presupposing something that is itself meaningful – namely timeless meaning. To be able to mean anything, one must already be able to draw on the proper vehicles for communicating that meaning. Intentions cannot do that work on their own.

How does this relate to Searle's criticism and his American soldier example? I think that Searle's objections and his example in particular bring out very clearly why we cannot have meaning reduce to intentions without relying on some form of conventional meaning. Of course, according to my analysis, this follows from the objections and the American soldier example in a roundabout way, since I agreed with Grice that the most natural way to describe the example is as a case of natural meaning. It is quite

[20] As Neale rightly points out, this does not undermine the idea that what an utterer means is determined by his communicative intentions. Cf. Neale (1992, 553). It does, however, put a rational constraint on what the (semantic) preconditions of communication are.

difficult to imagine how the American soldier should have brought his captors to believe that the *words* "*Kennst du das Land, wo die Zitronen blühen?*" mean "I am a German officer". But it is precisely this difficulty that forces us to deliberate *why* the American soldier cannot just get the soldiers to think that that is what those words mean. And the answer is that one cannot mean anything by anything whatsoever when hearer and speaker do not have a shared grasp of timeless meaning to fall back on, as happens when one is trying to speak to someone without having a shared language to mediate. Ironically, this is something made excessively clear by Grice's myth about the presumed origins of language, through which Grice inadvertently reveals how powerless (if conceivable at all) intentions are when there are no conventions to fall back on (Grice 1989c, 290-297).

The idea of autonomous intentions further recedes when one recalls how Grice himself describes the process of communication. We said that a hearer must in general grasp the literal meaning of an utterance to be able to decipher its speaker's meaning, and that the decision whether a given utterance is literal, as well as its re-interpretation in cases of non-literal speech, follows rational principles which were laid out in "Logic and Conversation". In addition to these principles and the utterance's general usage, a hearer intending to interpret an utterance can, according to Grice, (sometimes) rely on explicitly formulated linguistic (or quasilinguistic) intentions, the context of the utterance (linguistic or otherwise) or, in difficult cases, a deduction to determine the speaker's meaning (cf. Grice 1989b, 222-223; Neale 1992, sct. 6). Now it is striking that there is recourse to intentions only in the case of the explicitly formulated intentions, which are introduced with the caveat that they are *not* conclusive. According to Grice, "a speaker who has declared an intention to use a familiar expression in an unfamiliar way may slip into the familiar use" (Grice 1989b, 222). In other words, even when someone announces explicitly how to take his words, his deeds determine the outcome. Not only are intentions dependent on timeless meaning to even be articulated (let alone communicated). Even after having been formed *and* explicitly verbalized, timeless meaning still serves as an interpretive device that helps determine whether any given utterance was really intended in the way that the utterer has claimed it is. So that timeless meaning is both a presupposition for the functioning of intentional communication, and a retroactive corrective.

None of this is to say that intentions play no role in the process of communication, nor even that they play no role in constituting an utterance's meaning. Firstly, to ask what is meant by an utterance is to ask what a speaker *meant* by uttering a sentence, even when that sentence is intended by her to be taken literally. In other words, on a Gricean picture it is the speaker's intentions one is after when working out the meaning of an utterance U even in those cases where its vehicle is timeless meaning. Secondly, in working out the meaning of an utterance U that is not intended literally, one will only be able to understand the speaker by trying to figure out how she *meant* her utterance (which the hearer can do by taking the speaker to observe the Cooperative Principle and by applying the interpretation procedure sketched above). Thirdly, on a Gricean picture intentions play a role in the genesis of timeless meaning, insofar as their successful transfer once made it possible to develop a language and through it a means of communicating (and perhaps even forming) complex intentions.

As Neale and Strawson have pointed out, there is thus no circularity or inconsistency in Grice's original project (cf. Neale 1992, scts. 5 & 6; Strawson 1971, 174-175). Even if we concede that, as I have argued, intentions do not get off the ground without some conventions to fall back on, you can explain how intentions fit into the larger picture of the Gricean project without opening the project to definite objections.[21] Still I believe it is often overlooked that the consistency of Grice's project in "Meaning" comes at a price. If we follow Neale and Strawson in having (complex) communication rely on conventions, we distance ourselves from a view according to which intentions are what makes utterances meaningful in the first place. Whether we communicate via literal or nonliteral speech, timeless meaning is needed to work out (and perhaps even form) speaker's intentions. Thus, timeless meaning is a function of a community intending it to mean what it means only in the sense that they determine that individual utterance's meaning, but the fact that they can thus intend it to mean what it means is itself already a function of being able to intend utterances to mean something. If this is the right way to describe Grice's project in "Meaning", it is not concerned with explaining how meaning comes about, but with explaining how *given* our ability to mean things through utterances, individual

[21] See Avramides for a discussion of the advantages of a "weak, nonreductive interpretation of Grice's analysis" (Avramides 1989, 19). Avramides (1989, ch. 1).

utterances come to mean what they mean. In this sense, intentions are not constitutive of meaning because you cannot make sense of communication intentions independent of pre-existing meaning. The project of reducing timeless meaning to speaker's meaning without semantic remainder – which I have treated as the aim and defining feature of any intention-based semantics – cannot be pursued with the tools of "Meaning" alone.[22]

Strawson explicitly addresses this objection in "Meaning and Truth". He agrees that it would be absurd to credit ourselves with "extremely complicated communication-intentions (or at least desires)" independently of "linguistic means of fulfilling those desires" (Strawson 1971, 174). And he does seem to think that a project of the Gricean sort would falter should there be nothing more to say in its favor. That said, Strawson also believes that the project only requires that you can explain conventions of communication "in terms of the notion of pre-conventional communication at a rather basic level" (Strawson 1971, 174). And this is something he deems possible along the lines of Grice's already mentioned genetic account:

> Suppose an utterer achieves a pre-conventional communication success with a given audience by means of an utterance, say *x*. He has a complex intention, *vis-à-vis* the audience of the sort which counts as a communication-intention and succeeds in fulfilling that intention by uttering *x*. Let us suppose that the primary intention was such that the utterer *meant*

[22] Thus Neale's establishment of the non-circularity of Grice's project is orthogonal to my concerns. Neale seems to want to establish – pace objections that claim the contrary – "that typically the hearer must establish what *U* has said (or made as if to say) in order to establish what *U* meant; and it is by taking into account the nature and purpose of rational discourse that the hearer is able to progress (via, e.g., conversational implicature) from what *U* has said (or made as if to say) to what *U* meant" (Neale 1992, 552). It is precisely because I agree with Neale in this (see section 1 of this paper) that I see a problem for Grice. Neale does not acknowledge this issue because he appears to be concerned a) with a broadly Gricean approach (rather than "Meaning" taken in isolation), and because he seems to hold that b) Grice can get by without a strong, reductive approach. Since my paper is concerned with showing that (pace Searle) Grice's earlier approach is *not* a viable alternative to his later approaches, I treat "Meaning" in isolation, which no longer leaves open the possibility of holding on to a weak, reductive reading (since the project of "Meaning" builds on meaning fully reducing to intentional states).

> that *p* by uttering *x*; and, since, by hypothesis, he achieved a communication-success, he was so *understood* by his audience. Now if the same communication-problem presents itself later to the same utterer in relation to the same audience, the fact, known to both of them, that the utterer meant that *p* by uttering *x* before, gives the utterer a reason for uttering *x* again and the audience a reason for interpreting the utterance in the same way as before. (The reason which each has is the knowledge that the other has the knowledge which he has.) So it is easy to see how the utterance of *x* could become established as between this utterer and this audience as a means of meaning that *p*. Because it has worked, it becomes established; and then it works *because* it is established. And it is easy to see how this story could be told so as to involve not just a group of two, but a wider group. So we can have a movement from an utterer pre-conventionally meaning that *p* by an utterance of *x* to the utterance-type *x* conventionally meaning that *p* within a group and thence back to utterer-members of the group meaning that *p* by a token of the type, but now *in accordance with the conventions*. (Strawson 1971, 174-175)

This is not a *prima facie* implausible account. In fact, one could go so far as to claim that science gives us evidence of convention-fixing of the above sort, say among primates. The issue with this solution is rather that its plausibility is seriously strained when one fills in the specific details of Grice's account, which are of no concern to Strawson in "Meaning and Truth". Remember that to say that a speaker S meant something by X, according to Grice's original account that was rehearsed above, is to say that S intended the utterance of X to produce an effect in a hearer H (whether it is belief or understanding) by means of the recognition of this intention. For Strawson's argument to work as a defense of Grice's account, Strawson would be committed to claiming that the utterer in the above example is not only trying to make his hearers believe something by producing a cue (as in Grice's Herod example), but that he is trying to produce an effect in them by their recognition of his intention to that effect. This is a fairly "complex" intention to ascribe to a being with no linguistic means of fulfilling it. Does not Strawson fully embrace the feared absurdity when he credits utterers and hearers with no prior conventions of communication with being able to use their mutual "knowledge" as a "reason" for

"interpreting" utterances in the same way as before? While there are surely simple forms of communication that work along the broad lines sketched above, it is a stretch to couch them in such rationalistic terms (cf. Avramides 1989, 162-163). If we want to adopt a Gricean approach, communication is a deeply rational endeavor which is not instantiated by regularities among stimuli responses. Strawson's approach is just the first step in a long story about how we could get from stimuli to self-referential communication-intentions. Whatever happens when animals without language communicate, they do not communicate by wanting to get each other to understand or believe something in virtue of the recognition of that intention.[23]

The reason why the full force of the tension I have discussed has not been appreciated, seems to be that it is usually treated as a charge of inconsistency. Neale and Strawson focus on the question whether Grice can be reinterpreted consistently, that is whether his account rests on premises that undermine the project.[24] But the issue is not just whether there is a way of reinterpreting Grice consistently, but whether Grice's project can be reconstructed consistently while preserving its apparent aim. To my mind, it seems clear that Grice is interested in reducing meaning to intentions without semantic remainder rather than in merely explicating it in terms of the latter. One should keep in mind that timeless meaning, which is needed in any form of communication (even in nonliteral speech), is a function of a community of speakers intending it to mean what it means, which function

---

[23] It could be objected that this artificially creates a problem for Grice because I am here sticking to the self-reflexive intention which has in later works been dropped both by Grice himself and by most philosophers drawing on Grice to further the project in their own ways. Were it not for the self-reflexive intention, the problem would not seem to persist. Having said that, my paper is intended to problematize the original account as presented in "Meaning" and is thus orthogonal to the issue of, say, the late Grice/Schiffer amendments. My aim has been to show that Searle's calls to save the Gricean account by returning to "Meaning" and its self-reflexive intention are to be rejected.

[24] Neale's main concern in this regard is to show that Grice can a) account for the way in which conventional meaning plays a role in working out communicative intentions (cf. section 1 of this paper), and b) explain how the meaning of a sentence is (partly) determined by its parts. For these reasons, Neale rejects the view that Grice's account is either circular or absurd. See Neale (1992, 544, 550-552).

is itself a function of the ability of intending utterances to mean something. In this sense, Grice's account is not really an example of an intention-based semantics, as intentions must be conceived as carriers rather than constituents of meaning. Grice's original account can explain why utterances mean what they mean (rather than meaning something else), but it cannot explain how it comes about that we can even mean things by sharing utterances.

3. Conclusion

The analysis of Searle's American soldier example was meant to bring out, from the perspective of Grice's original account from "Meaning", the need and simultaneous inadequacy of having utterances rely on a shared repository of conventional meaning. If we postulate the reliance of intentions on prior meaning, we give up any hope of reducing timeless meaning to speaker's meaning without semantic remainder, and with it the project of an intention-based semantics. If we do not, we cannot explain how we can come to form or communicate complex intentions precisely because we have no rules or conventions to fall back on. The project of "Meaning" thus fails as an attempt at constructing an intention-based semantics: it is not possible to analyze timeless meaning in terms of speaker's meaning without semantic remainder, as speaker's meaning is itself reliant on prior timeless meaning. As I have tried to show, this is a serious blow to any attempts (even non-reductionist ones) at rehabilitating Grice's original account, which does indeed seem committed both to self-reflexive intentions and full analyzability of timeless meaning in terms of speaker's meaning.

Acknowledgments

I would like to thank Clemens Schmalhorst, Thomas Enthofer, Niklas Ernst, Jonas Hertel, and an anonymous referee for their helpful remarks and criticisms.

References

ARMSTRONG, D. M. (1971): Meaning and Communication. *The Philosophical Review* 80(4), 427-447.

AVRAMIDES, A. (1989): *Meaning and Mind. And Examination of a Gricean Account of Language*. Cambridge MA: MIT Press.

BENNETT, J. (1973): The Meaning-Nominalist Strategy. *Foundations of Language* 10(1), 141-168.

BRANDOM, R. B. (1994): *Making It Explicit. Reasoning, Representing, and Discursive Commitment*. Cambridge MA: Harvard University Press.

BURGE, T. (1979): Individualism and the Mental. *Midwest Studies in Philosophy* 4, 73-121.

GRANDY, R. & WARNER, R. (2017): Paul Grice. In: Zalta, E. N. (ed.): *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), <https://plato.stanford.edu/archives/win2017/entries/grice/>.

GRICE, P. (1989a): Logic and Conversation. In: Grice, P.: *Studies in the Way of Words*. Cambridge MA: Harvard University Press, 22-40.

GRICE, P. (1989b): Meaning. In: Grice, P.: *Studies in the Way of Words*. Cambridge MA: Harvard University Press, 213-223.

GRICE, P. (1989c): Meaning Revisited. In: Grice, P.: *Studies in the Way of Words*. Cambridge MA: Harvard University Press, 283-303.

GRICE, P. (1989d): Retrospective Epilogue. In: Grice, P.: *Studies in the Way of Words*. Cambridge MA: Harvard University Press, 339-385.

GRICE, P. (1989e): Utterer's Meaning, Sentence-Meaning, and Word-Meaning. In: Grice, P.: *Studies in the Way of Words*. Cambridge MA: Harvard University Press, 117-137.

GRICE, P. (1989f): Utterer's Meaning and Intentions. In: Grice, P.: *Studies in the Way of Words*. Cambridge MA: Harvard University Press, 86-116.

KRIPKE, S. A. (1982): *Wittgenstein on Rules and Private Language*. Cambridge MA: Harvard University Press.

LYCAN, W. G. (2008): *Philosophy of Language. A Contemporary Introduction*. 2. Ed. NY/London: Routledge.

MARTINICH, A. P. (1984): *Communication and Reference*. Berlin/NY: Walter de Gruyter.

NEALE, S. (1992): Paul Grice and the Philosophy of Language. *Linguistics and Philosophy* 15(5), 509-559.

RECANATI, F. (1986): On Defining Communicative Intentions. *Mind & Language* 1(3), 213-242.

SCHIFFER, S. (1972): *Meaning*. Oxford: Clarendon Press.

SEARLE, J. R. (1969): *Speech Acts. An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

SEARLE, J. (2007): Grice on Meaning: 50 Years Later. *Teorema* 36(2), 9-18.

SPERBER, D. & WILSON, D. (1986): *Relevance. Communication and Cognition*. Cambridge MA: Harvard University Press.

STRAWSON, P. F. (1971): *Meaning and Truth*. Reprinted in: P. F. Strawson: *Logico-Linguistic Papers*. London: Methuen, 170-189.

VLACH, F. (1981): Speaker's Meaning. *Linguistics and Philosophy* 4(3), 359-391.

YU, P. (1979): On the Gricean Program about Meaning. *Linguistics and Philosophy* 3, 273-288.

ZIFF, P. (1967): On H. P. Grice's Account of Meaning. *Analysis* 28, 1-8.

Organon F 25 (4) 2018: 589-590

Modal Metaphysics: Issues on the (Im)Possible VI
Bratislava, August 2-3, 2018[1]

Modal Metaphysics: Issues on the (Im)Possible conference has commenced its existence in 2013. This year, in 2018, the conference brought its 6th instalment aiming basically at the same thing as at the beginning: to overview the current research on modality. Be it metaphysics, epistemology, formal logic, semantics or fiction, all the presented papers proved the increasing interest in the field.

The conference kicked-off with two talks: Gaétan Bovey's (University of Neuchâtel, Switzerland) "Can 'Intrinsicality' Save the Existential-modal Account of Essence? A Critical Response to David Denby" commented by Karol Lenart, and Michael J. Raven's (University of Victoria, Canada & University of Washington, USA) "A Problem for Immanent Universals in States of Affairs" followed by Riccardo Baratella's comments. Daniel Milne-Plückebaum (Bielefeld University, Germany) then proposed "Meinongian Modal Meinongianism" and Matthew James Collier (University of Oxford) presented a paper "God Exists in all Possible Worlds: Anselmian Theism and Genuine Modal Realism" (commented by Daniel Berntson). The co-authored paper by Anand Jayprakash Vaidya (San José State University, USA) and Michael Wallner (University of Graz, Austria) motivated "Reductive and Non-Reductive Finean Essentialism" (commented by Gaétan Bovey) and Giacomo Giannini (Durham University, UK), followed by Sanna Mattila's reaction, approached "Resemblance, Representation, and Counterparts" trial. After it, Matthew James Collier discussed "Impossible authorships? Or how could Pierre Menard be the author of The Quixote" by Jorge Luis Méndez-Martínez (National Research University in Moscow, Russian Federation) and Sanna Mattila's (University of Helsinki, Finland) "Epistemology of Possibility and Reliabilism: a Challenge Considered" received comments from David Mark Kovacs. The last dual of talks were delivered by Michael De (University of Miami, USA) and Nathan Hawkins (Cambridge University, UK), commented by Michael Wallner and Matteo Pascucci, respectively. The end of the first

[1] ✉ Martin Vacek

Institute of Philosophy
Slovak Academy of Sciences
Klemensova 19, 811 09 Bratislava, Slovakia

e-mail: martinvacekphilosophy@gmail.com

day belonged to Gonzalo Rodriguez-Pereyra. His keynote lecture "Why is there Something Rather than Nothing? A Probabilistic Answer Examined" both presented the original Peter van Inwagen's answer to the question 'Why is there something rather than nothing?' and challenged his argument by challenging two of its premises.

The second day of the conference started also with two parallel sessions: Fernando Furtado's (University of Lisbon, Portugal) "S5- denying Approach to Relativised Metaphysical Modality" (commentated by Nathan Hawkins) and Daniel Berntson's (Princeton University, USA) "Relational Possibility". David Mark Kovacs (Tel Aviv University, Israel) delivered a paper entitled "Constitution, Dependence, and Mereological Hylomorphism" followed by Jorge Luis Méndez-Martínez's comments and Giacomo Giannini commented on Riccardo Baratella's (University of Padua, Italy) "Material Objects, Events, and Property Instances". Karol Lenart (Jagiellonian University, Poland) with (Michael De assigned as his commentator) overviewed "Essentialism, Haecceitism and AntiHaecceitism" while Daniela Glavaničová and Miloš Kosterec reviewed Bjørn Jespersen's (VSB-TU Ostrava, Czech Republic and University of Utrecht, Netherlands) "The Man without Properties: Impossible Individuals as Hyperintensions" contribution. The accepted talks ended up with Moritz Baron's (The Universities of Stirling and St Andrews, Scotland) "Can Williamson's Counterfactual-based Epistemology of Modality Explain our Knowledge of Mathematical Necessity?" (with Michael J. Raven as a commentator) and Cristina Nencha's (University of Turin, Italy) "David Lewis and Kit Fine's Essences". The end of the conference fulfilled the second keynote lecture give by Sonia Roca-Royes. Roca-Royes explored the prospects of rationalist, concept-based epistemologies of modality and concluded that concepts have at most a limited role to play in the epistemology of essence (and de re modality).

For the first time the conference has a younger tense counterpart: Truth in Time and Open Future stream. The stream hosted five talks: Giacomo Andreoletti (University of Tyumen, Russian Federation): "Time Travel, Freedom, and Branching Time"; Michael De (University of Miami, USA): "The Open Future and Likelihood"; Vincent Grandjean (University of Neuchâtel, Switzerland): "How is the Asymmetry between the Open Future and the Fixed Past to be characterized?"; Tomáš Kollárik (Commenius University in Bratislava, Slovakia): "The Assertion Problem" and Elton Marques (University of Lisbon, Portugal): "Determinism, Eternalism and the Stheory". Idle to say, we always gladly welcome all the contributions from all parts of the world. We do so by following a basic rule of any conference: a conference is as good as its participants are. This report verifies the validity of the rule and, hopefully, the next report will do the same.

Martin Vacek

Modal Metaphysics: Issues on the (Im)Possible VII

May 30-31, 2019
(Bratislava, SLOVAKIA)

Keynote speakers

MATTHEW BRAHAM (University of Hamburg, Germany)

GREGORY CURRIE (University of York, UK)

PETER VAN INWAGEN (University of Notre Dame, USA)

(more keynote speakers will be announced soon)

We invite submissions for a 30 minute presentation followed by 10 minute comments and a 15 minute discussion. Areas of interest include any aspect of analytic metaphysics, epistemology and logic of modality.

A paper of approximately 3000 words should be prepared for blind review and include a cover page with the full name, title, institution and contact information. Papers can be submitted in pdf or doc(x) and should be sent to **modalmetaphysics@gmail.com**. Talks will be followed by commentaries.

Deadline for submission: January 15, 2019

Notification of acceptance: February 28, 2019

Contact:
Martin Vacek (martinvacekphilosophy@gmail.com)

Conference website: www.metaphysics.sk.

Modal Metaphysics: Issues on the (Im)Possible VII

(cont.)

May 30-31, 2019
(Bratislava, SLOVAKIA)

As a part of the conference, we will host three additional workshops:

Current Trends in Deontic Logics II: Modality, Hyperintensionality, Responsibility

Philosophy of Language: Semantics of Fictional Discourse II

Truth in Time and Open Future II

For these workshops, we invite submissions for 30 minutes talk and subsequent 15 minutes discussion. An abstract with 400 – 600 words including full name, title, institution, and contact information should be prepared for blind review and sent to **modalmetaphysics@gmail.com**. Please, put the name of the relevant workshop in the subject line.

Deadline for submission: January 15, 2019

Notification of acceptance: February 28, 2019

Contacts:

Current Trends in Deontic Logics II: Daniela Glavaničová (daniela.glavanicova@gmail.com) or Matteo Pascucci (matteo.pascucci@tuwien.ac.at)

Semantics of Fictional Discourse II: Martin Vacek (martinvacekphilosophy@gmail.com)

Truth in Time and Open Future II: Tomáš Kollárik (mimoine@gmail.com)

Contents

ARTICLES

DISCUSSIONS

BOOK REVIEWS

REPORTS