Contents

ARTICLES

DISCUSSIONS

BOOK REVIEWS

Organon F 25 (2) 2018: 142-154

Kripkean Essentialist Argument and Its Generalization

JAEHO LEE[1]

ABSTRACT: In this paper I examine the argument by H. Beebee and N. Sabbarton-Leary that Brian Ellis's scientific essentialism is based on the "abuse" of the necessary *a posteriori*. I will first briefly survey various attempts to resist what I will call the "Kripkean essentialist argument" to locate Beebee's and Sabbarton-Leary's position properly. After that I will argue that Beebee's and Sabbarton-Leary's argument is not successful; in particular, I will argue that under the most natural interpretation of their position it is not internally coherent, and that their argument is based on a superficial understanding of Kripkean necessity *a posteriori*.

KEYWORDS: Analyticity – Kripkean essentialist argument – natural kind – necessity a posteriori.

1. Kripkean essentialist argument

In this paper I will use "the Kripkean essentialist argument" (KE) as an overarching term that embraces both "the general version of Kripkean essentialist argument" (GKE) and "the special version of Kripkean essentialist argument" (SKE). GKE has the following components.

[1] Received: 21 July 2017 / Accepted: 30 October 2017

✉ Jaeho Lee

Department of Philosophy, Chung-Ang University
84 Heukseok-ro, Donjak-gu
156-756 Seoul, Korea

e-mail: jaeho.jaeho@gmail.com

(1) "Water is H_2O" is *a posteriori*.

(2) "Water is H_2O" is necessary.

(3) "Water is H_2O" is necessary *a posteriori*. [from (1), (2)]

(4) Water is essentially H_2O.[2]

(5) This (type of) argument can be applied to all natural kinds.

The role of (5) here is to generalize the argument expressed in (1) – (4). This embedded argument is what I will call "the special version of the Kripkean essentialist argument" (SKE).

Philosophers have resisted KE in various ways. Some philosophers deny (1). For example, J. LaPorte argues that "Water is H_2O" is necessary but not *a posteriori* (see LaPorte 2004). According to him, when scientists discovered the chemical constitution of water, they stipulated thereby that water is H_2O. Therefore, "Water is H_2O" is indeed necessary; however, since this necessity comes from the stipulation, it is not *a posteriori*.

Other philosophers cast doubt on (2). For example, some experimental philosophers think that the Kripke/Putnam-style intuition beyond (2) is dubious (see Machery et al. 2004 and Weinberg 2007). Since, *pace* LaPorte, "Water is H_2O" is not analytic, its necessity should be shown by something like Putnam's Twin Earth argument and these philosophers argue that the anti-descriptivist intuition appealed to in Putnam-style arguments is significantly weak among East Asians, which casts doubt on the reliability of this intuition.

Still other philosophers deny that (4) follows from (3) (or that (4) is equivalent to (2)). In other words, these philosophers think that Kripkean necessity *a posteriori* has no metaphysical implications. For example, Alan Sidelle argues that although we should accept that there is such a thing as necessity *a posteriori*, Kripkean necessity *a posteriori* is a mere consequence of linguistic convention or linguistic decision (see Sidelle 2002, 310). According to him, "Water is H_2O" is necessary *a posteriori* simply because we have agreed collectively to use "water" as a rigid designator.

---

[2] There might be different justifications concerning (4). One might think that (4) follows from (3). Others might think that (4) follows directly from (2), because what Kripke means by "necessity" in (2) is a metaphysical one. I think that this difference does not make any big difference in my arguments below.

This is a mere linguistic decision. If we had decided to use "water" as a descriptor, "water is H_2O" would not have been necessary. Since linguistic decisions do not change the world, Kripkean necessity *a posteriori* has no metaphysical implications, which means that we cannot infer (4) from (3) (or from (2)).

Unlike these arguments, the argument by H. Beebee and N. Sabbarton-Leary (henceforth BS),[3] which is the main topic of this paper, seems to focus on (5). They say, "Even if we accept that Kripke's story holds for proper names and natural kind terms, it can by no means be taken for granted that the story extends to cover other cases. This paper rehearses the general argument that such arguments are indeed required, and discusses in detail one examples of abuse of the necessary *a posteriori*: Brian Ellis's 'scientific essentialism'" (Beebee & Sabbarton-Leary 2010, 159).[4] Given this, it is natural to think that BS have no explicit objection to SKE. If this is correct, their argument is intended to be distinct from the above three types of arguments in terms of its target.[5]

2. BS's argument

BS's main example for their claim that (5) is false is that of ununbium. "Ununbium" is a temporary designator for element 112, which was first discovered (created) by Sigurd Hoffman and his team in the mid-1990s (and has now been formally recognized by the International Union of Pure and Applied Chemistry (IUPAC) and given the permanent name Copernicium). According to the standard system for temporary naming used by

---

[3] See Beebee & Sabbarton-Leary (2010). For Ellis's scientific essentialism, see Ellis (2001; 2002).

[4] Here what BS mean by "other cases" is the cases of such natural kinds as ununbium which do not have Kripkean natural kind term.

[5] BS might claim that my interpretation misrepresents their intention. All they want to say is, they might claim, that Ellis needs an argument for (5) and that he failed to provide one. Under the current context, the correctness of my interpretation is not very important. If what BS want to show is that (5) needs an argument rather than that (5) is false, then my criticism of BS in this paper can be regarded as an argument BS requires.

IUPAC, element 112 becomes "ununbium": un(1) + un(1) + bi(2) + (i)um. From this example, BS conclude as follows.

> [What the example of ununbium illustrates is] that some – and indeed clearly most – chemical names are not introduced using a Kripke-style name-acquiring transaction. Rather, they are generated using a complex set of rules and grammar, and clearly encode descriptive information. In other words, they are descriptors. As a result, a theoretical identity sentence such as 'ununbium is the element with atomic number 112' […] is something a chemist can come to know *a priori*. (Beebee & Sabbarton-Leary 2010, 165)

If this conclusion is correct, as BS argue, (5) is false. One might think, however, that the mere fact that (5) is false does not by itself undermine Ellis' scientific essentialism because it does not follow from the fact that "ununbium is the element with atomic number 112" is *a priori* (or analytic) that it is not the case that ununbium is essentially the element with atomic number 112.[6] After all, what Ellis really wants to show is that his scientific essentialism is true rather than that "ununbium is the element with atomic number 112" is necessary *a posteriori*. Even if the latter turns out to be false, as long as his scientific essentialism is intact, the situation is not very painful for Ellis.

However, BS argue that the situation is much worse than this because, given that GKE does not work, there is no way for Ellis to show that ununbium is essentially the element with atomic number 112.[7] So the falsity of (5) has the consequence that we have no good reason to accept Ellis's scientific essentialism either.

---

[6] I am not saying that this is what Ellis actually thinks. BS claim that Ellis "is committed to the view that analytic truths cannot be truths about essences" (Beebee & Sabbarton-Leary 2010, 173).

[7] In fact, the story is much more complicated than this. Ellis does provide his own criterion for distinguishing analytic necessity from metaphysical necessity and it does not directly appeal to GKE. But BS convincingly argue that this criterion does not work (Beebee & Sabbarton-Leary 2010, 173-174).

3. Two problems with BS's argument

I think there are at least two problems with BS's argument. The first is that BS's position is extremely unstable and its internal coherence is dubious. Let me assume that the main target of BS's argument is (5) and that they have no explicit objection to (1) – (4); as I said before, this is the most natural interpretation of their position. Let me assume further that their argument is successful. Then it seems that they should say something like this.

(6) While (as Kripke claims) gold is essentially the element with atomic number 79, it is not the case (or at least there is no reason to think) that ununbium is essentially the element with atomic number 112.

"Gold" is similar to such a proper name as "Nixon" in that it is non-descriptive and rigid. So there is no problem with applying the Kripke-style argument we find in (1) – (4) to "gold". But unlike "gold," according to BS, "ununbium" is a descriptor. In this case, no Kripke-style argument is applicable to "ununbium". Given what BS say, there is no other way to show that ununbium is essentially the element with atomic number 112. So we are left with (6).

This is a weird conclusion. If we can say that gold is essentially the element with atomic number 79, why is it not allowed to say that ununbium is essentially the element with atomic number 112? The lack of homogeneity in the metaphysical picture this conclusion implies is extremely unsatisfactory and should be avoided, if possible. There seem to be two potential ways to avoid it. The first is to use some kind of inductive generalization. We know that gold has its atomic number essentially. We know iron has its atomic number essentially. We know copper has its atomic number essentially. So, we have an inductive generalization: all elements have their atomic numbers essentially. Since ununbium is an element, and its atomic number is 112, ununbium is essentially the element with atomic number 112. In short, we can show that ununbium is essentially the element with the atomic number 112 without applying the Kripke-style argument directly to "ununbium". If this is correct, BS's claim that there is no other way to show that ununbium is essentially the element with atomic number

112 is wrong. Since what Ellis really wants to show is that scientific essentialism is true rather than that (5) is true, BS's argument that (5) is false is not particularly painful for Ellis.

At this point, BS might argue that the above inductive generalization is not justified. They might claim that while "gold", "iron", and "copper" are all introduced using a Kripke-style name-acquiring transaction, "ununbium" is introduced in a completely different way, and that this difference blocks the inductive generalization. They might go on to say that in such a case, the only justifiable inductive generalization is that all elements *whose name is introduced using a Kripke-style name-acquiring transaction* have their atomic numbers essentially. This is the point where the second way to avoid the lack of homogeneity in our metaphysical picture comes into our story. To say that this is the only justifiable generalization is to say that the way the name of an element is introduced is critical in deciding whether something similar to SKE is applicable to that element. However, the way a name is introduced does not change the world. To make this clear, consider the following.

(7) Ununbium does not have its atomic number essentially. But if the name of ununbium had been introduced using a Kripke-style name-acquiring transaction, then ununbium would have its atomic number essentially.

Obviously (7) is not acceptable. Given this, the best thing BS can do is to say that Kripkean necessity *a posteriori* has no metaphysical implications. In this case, we are not allowed to infer (4) from (3) (or from (2)), which means that BS's argument is not very different from Sidelle's argument explained above. In other words, in this case, contrary to appearances, the main target of BS's argument is not (5) but the inference (4) from (3) (or from (2)).

To summarize, BS need to clarify their position, and it seem that they have three options. The first is to embrace a nonhomogeneous metaphysical picture: Gold is essentially the element with the atomic number 79 but ununbium is not essentially the element with the atomic number 112. The second is to say that their argument that (5) is false has no relevant metaphysical implications and that Ellis' scientific essentialism is still tenable. The third is to say that contrary to appearances, what their argument shows

is something very similar to Sidelle's claim that Kripkean necessity *a posteriori* is just a linguistic phenomenon with no metaphysical implications.

I believe that none of them is satisfactory to BS. The metaphysical picture the first option requires is too weird to accept when there are other pictures that do not have such a consequence. The second option deprives BS's argument of its teeth: Their argument might be sound. But it does not undermine Ellis' scientific essentialism. The third option is not satisfactory either, because it makes BS's argument into a not particularly novel one that merely pretends to novelty.

The second problem, which I find more serious, is that it is not clear whether BS's argument succeeds in showing that (5) is false. Even if it does succeed in showing this, I am pretty sure that it cannot show that the following variation of (5) is false.

(5′) This argument, *or something very similar to this argument*, can be applied to all natural kinds.

Here is my argument. First imagine an Earth-like planet (call it "U-Earth") where ununbium is as abundant as water on our Earth. In addition, imagine that the people who live on that planet call ununbium "unux" and that this name is introduced using a Kripke-style name-acquiring transaction. Now we can make the following argument.

(1′) "Unux is the element with atomic number 112" is *a posteriori* on U-Earth.

(2′) "Unux is the element with atomic number 112" is necessary on U-Earth.

(3′) "Unux is the element with atomic number 112" is necessary *a posteriori* on U-Earth. [from (1′), (2′)]

(4′) Unux is essentially the element with atomic number 112 on U-Earth. [from (3′)]

(4″) Ununbium is essentially the element with atomic number 112 *on our Earth*. [from (4′)[8]]

[8] The inference (4″) from (4′) is based on the assumption that accessibility relation between possible worlds is transitive. Some philosophers, for example N. Salmon, deny

This argument is very similar to SKE and seems to appeal to the same intuition. But, unlike SKE, this argument works even if "ununbium" is *not* introduced using a Kripke-style name-acquiring transaction. Therefore, even if BS's argument is successful in showing that (5) is false, it cannot show that (5′) is false.

BS might claim that this argument does not work because it uses U-Earth English in (1′) – (4′) but uses our English in (4″). However, I think that (1′) – (3′) are clearly sentences in our English. Of course, these sentences contain a U-Earth English sentence, namely "Unux is the element with atomic number 112". But this sentence is not used but mentioned. What is problematic is (4′). I concede that it is natural to think (4′) is a U-Earth English sentence, but I believe that this does not make any difference. If we want to be consistent, we may use the following instead of (4′).

(4‴) "Unux is essentially the element with the atomic number 112" is true on U-Earth.

(4‴) is clearly a sentence in our English. And we can infer (4″) from (4‴): If we know that "Wasser ist im Wesentlichen H_2O" is true in German, under the assumption that we understand this German sentence, we can safely conclude that water is essentially H_2O.

There is another argument that need not deal with this kind of complexity. It goes like this.

(a) If the name of ununbium had been "Unux" and it had been introduced using a Kripke-style name-acquiring transaction, "Unux is the element with atomic number 112" would be *a posteriori*.

(b) If the name of ununbium had been "Unux" and it had been introduced using a Kripke-style name-acquiring transaction, "Unux is the element with atomic number 112" would be necessary.

(c) If the name of ununbium had been "Unux" and it had been introduced using a Kripke-style name-acquiring transaction, "Unux

---

this assumption. These philosophers might think that although it is possible that Ununbium is essentially the element with atomic number 112, it does not follow from that that Ununbium is essentially the element with atomic number 112. For Salmon's view and its problem see Roca-Royes (2016).

is the element with atomic number 112" would be necessary *a posteriori*. [from (a), (b)]

(d) If the name of ununbium had been "Unux" and it had been introduced using a Kripke-style name-acquiring transaction, "Unux is essentially the element with atomic number 112" would be true.

(e) The way a name is introduced does not change the world.

(f) Therefore, ununbium is essentially the element with the atomic number 112. [from (d), (e)]

Note that every sentence used in this argument is in our English.

If at least one of these two arguments is sound, BS's argument once again loses its teeth. Their argument might be able to show that (5) is false, but it does not show that (5′) is false; in fact, we have good reason to think (5′) is true. If so, there is no reason to think that BS's argument undermines Ellis' scientific essentialism.

4. Are analyticity and necessity *a posteriori* mutually exclusive?

The main idea behind BS's argument seems to be this.

(8) Analyticity and necessity *a posteriori* are mutually exclusive.

(9) "Ununbium is the element with atomic number 112" is analytic.

(10) So, this cannot be necessary *a posteriori*. [from (8), (9)]

(11) So, (5) is false.

(12) There is no other way for Ellis to justify his scientific essentialism.

(13) So, Ellis' scientific essentialism is not justified.

I have already argued that (12) is false, since Ellis can use (5′) instead of (5). In this section I will argue that (8) (and hence (10)) cannot be taken for granted. This is an important issue for both BS and Ellis. BS says "The first point that needs to be made about Ellis's position is that he simply takes it for granted that it is '*a posteriori* what properties are essential to a given

kind'" (Beebee & Sabbarton-Leary 2010, 163). The primary target of BS's ununbium example is Ellis's claim that all essence talk concerning natural kinds are necessary *a posteriori*.[9] If (8) (and hence (10)) is true, then this claim by Ellis must be false. I already said this result is not very problematic for Ellis as long as he can reject (13); but it is still problematic for him to some extent.

If my arguments in the previous section are correct, we may have a pretty good sense of a "necessity *a posteriori"* in which (8) is false. This seems to imply that we may have more than one senses of "necessity *a posteriori"*. Compare the following two definitions of "necessity *a posteriori*".

(Def1) S is necessary *a posteriori* iff S is necessary and its truth can be known *only* a *posteriori*.

(Def2) S is necessary *a posteriori* iff S is necessary and its *necessary* truth can be known *genuinely a posteriori*.

Some clarifications of (Def2) are in need. First, note that the presence of "only" is not the only difference between (Def1) and (Def2). In (Def2), we have "necessary truth" rather than "truth". This difference is important. If we use "truth" in (Def2), (Def2) becomes deeply unsatisfactory. Consider "all bachelors are unmarried men". The truth of this sentence can be known through an *a posteriori* method. Just examine a sample of bachelors and inductively generalize the observed regularity! In this light, we should say that this sentence is necessary *a posteriori*, which is absurd. However, our (Def2) does not have this problem. There is no (genuinely) *a posteriori* way to show the *necessary* truth of this sentence. Second, I need to explain why "genuinely" is required. Without this, one might think, even mathematical truths may become *necessary a posteriori*. Assume that S is a notorious mathematical proposition. Imagine that a famous mathematician finally proved S and that I read this in a newspaper. Now I know that S is true. But my knowledge seems to be *a posteriori*. This worry can be handled, however, if we insert "genuinely" in (Def2) and define this term as follows: The truth of a sentence is known *genuinely a posteriori* iff this knowledge is acquired *a posteriori* and it is not transmitted from someone

[9] This is why the title of their paper is "On the Abuse of the Necessary *a Posteriori*".

else's *a priori* knowledge of the truth of the sentence. Since my knowledge of S in the above example is transmitted from the *a priori* knowledge of the famous mathematician, it cannot be genuinely *a posteriori*.

Now, if we accept (Def1) what BS say is right: "Ununbium is the element with the atomic number 112" is not necessary *a posteriori*. If "ununbium" is a descriptor and hence this sentence is analytic, there is an *a priori* way to show the truth of this sentence, and it is automatically disqualified as a necessary *a posteriori* sentence. But if we accept (Def2), the story becomes completely different. Imagine that there is a chemist who is completely ignorant of the standard naming system of IUPAC but is familiar with the semantics and metaphysics of natural kind. He will not know whether "ununbium" is a descriptor or not, and so, he cannot know whether "ununbium is the element with atomic number 112" is analytic or not. But he can examine some samples of ununbium and find that ununbium is an element and that its atomic number is 112. After that, he can say like the following. I don't know whether the name of ununbium is introduced using a Kripke-style name-acquiring transaction, but I do know that if it had been introduced using a Kripke-style name-acquiring transaction, "Ununbium is the element with atomic number 112" would be necessary *a posteriori*. From this, I can know that ununbium is essentially the element with the atomic number 112. So, I can know that "Ununbium is the element with atomic number 112" is necessary. The method this chemist used is an *a posteriori* method. So, this story shows that even if "ununbium is the element with atomic number 112" is an analytic sentence, its necessary truth can be known *a posteriori*.[10] If we accept (Def2), this sentence is both analytic and necessary *a posteriori.* Thus, analyticity and necessity *a posteriori* are not mutually exclusive. A consequence of this is that Ellis's claim that all natural kinds produce necessity *a posteriori* is still tenable in the face of BS's ununbium example.

[10] An anonymous reviewer claimed that the chemist in my story uses "ununbium" in a different sense because she re-baptized ununbium. I disagree. It is quite uncontroversial that she *inherited* the name "ununbium" from other people in the Kripkean sense. One can inherit a name without knowing its etymology and Kripkean inheritance of name does not require this kind of knowledge either.

I believe that (Def2) is a pretty good way to define "necessity *a posteriori*" and that we cannot take it for granted that (Def1) is the right definition. This does not mean that there is no problem with (Def2). To be sure, it is not perfectly consistent with the conventional use of "a posteriori". For example, when we say "water is H_2O" is a posteriori, what we usually mean is that its truth can *only* be known *a posteriori*. Nevertheless, I think that there is a non-negligible motivation for (Def2) because, given my arguments in the previous section, (Def2) carves the joint of nature better than (Def1).[11] Once we accept (Def2), "necessity a posteriori" can subsume all essentialist claims about natural kinds. But if we accept (Def1), "necessity a posteriori" can subsume only a small part of these claims. For this reason, I think that BS were too quick in accepting (8).[12]

5. Conclusion

I think that many philosophers have underestimated the force of KE. As BS's case explicitly shows, philosophers have viewed the *actual* history of naming as crucial in KE. If this is true, I believe, KE is vulnerable to the Sidelle-style criticism that Kripkean necessity *a posteriori* is a mere linguistic phenomenon. However, if my arguments in this paper are correct, actual history of naming is not that important in KE at least as long as we accept SKE. I concede that it is important in some cases. As Kripke has shown plausibly, it is indeed important in the truth of such sentence as "Gödel was born in 1906" in the situation where not Gödel but Schmitt was the person who proved the incompleteness of arithmetic. However, once we assume that "ununbium" refers ununbium somehow, it is not important at all for the truth of "ununbium is essentially the element with atomic number 112" how this name is introduced. I believe that a moral we can

---

[11] For the importance of "carving the joint of nature" in interpretation, see Sider (2011, 23-35) and Lewis (1983).

[12] In fact, as BS point out, Ellis himself seems to accept (8) too (Beebee & Sabbarton-Leary 2010, 173). So BS could say that the criticism should apply not to them but to Ellis. They could say that all they wanted to show is that Ellis' position is not internally coherent. Here I am not very interested in the question of who should be blamed; rather, I am more interested in whether (8) is true.

learn from the failure of BS's argument is that we should not conflate the cases where we are talking about reference and the cases where we are talking about essence.

References

BEEBEE, H. & SABBARTON-LEARY, N. (2010): On the Abuse of the Necessary A Posteriori. In: Beebee, H. & Sabbarton-Leary, N. (eds.): *The Semantics and Metaphysics of Natural Kinds*. London: Routledge, 159-178.

ELLIS, B. D. (2001): *Scientific Essentialism*. Cambridge, U.K. & New York: Cambridge University Press.

ELLIS, B. D. (2002): *The Philosophy of Nature: A Guide to the New Essentialism*. Chesham: Acumen.

LAPORTE, J. (2004): *Natural Kinds and Conceptual Change*. Cambridge, U.K. & New York: Cambridge University Press.

LEWIS, D. (1983): New Work for a Theory of Universals. *Australasian Journal of Philosophy* 61(4), 343-377.

MACHERY, E., MALLON, R., NICHOLS, S., & STICH, S. P. (2004): Semantics, cross-cultural style. *Cognition* 92, B1-B12.

ROCA-ROYES, S. (2016): Rethinking Origin Essentialism (for Artefacts). In: Jago, M. (ed.): *Reality Making*. Oxford: Oxford University Press, 152-176.

SIDELLE, A. (2002): On the Metaphysical Contingency of Laws of Nature. In: Gendler, T. & Hawthorne, J. (eds.): *Conceivability and Possibility*. Oxford & New York: Clarendon Press, 309-336.

SIDER, T. (2011): *Writing the Book of the World*. Oxford & New York: Clarendon Press.

WEINBERG, J. M. (2007): How to Challenge Intuitions Empirically Without Risking Skepticism. *Midwest Studies in Philosophy* 31, 318-343.

Organon F 25 (2) 2018: 155-173

Truths, Facts, and Liars

PETER MARTON[1]

ABSTRACT: A Moderate Anti-realist (MAR) approach to truth and meaning, built around the concept of *knowability*, will be introduced and argued for in this essay. Our starting point will be the two fundamental anti-realists principles that claim that neither truth nor meaning can outstrip knowability and our focus will be on the challenge of adequately formalizing these principles and incorporating them into a formal theory. Accordingly, I will introduce a MAR truth operator that is built on a distinction between being true and being factual. I will show then that this approach partitions propositions into eight classes, on the basis of their knowability. We will then ask the following question: Given the anti-realist principles, what kind of theory of propositional meaning can properly explain the *meaninglessness* of *fully unknowable* propositions? This question will lead us to the claim that the *meaning/content* of propositions should be identified not with the set of possible worlds in which the propositions are *true/factual*, but rather in which they are *known*. This modified approach will then be used to analyze both the *Liar Paradox* and the *Strengthened Liar*. To anticipate the conclusion of this essay, it will be shown that a MAR framework can render definite truth and factuality values to the *Liar sentence* and it will also confirm our intuition that such paradoxical sentences are devoid of proper meaning.

KEYWORDS: Chuch-Fitch paradox – knowability – Liar Paradox – meaning – moderate anti-realism – truth.

[1] Received: 30 November 2017 / Accepted: 23 April 2018

✉ Peter Marton

Department of Philosophy, Bridgewater State University
131 Summer Street,
Bridgewater, MA 02325, U.S.A.

e-mail: pmarton@bridgew.edu

0. Introduction

One standard way of approaching a certain class of semantic paradoxes (as e.g. the *Liar*, the *Knower*, etc.) is to claim that the crucial sentence in the setup of the paradox is *meaningless*. This approach is not without problems: first, the crucial sentences in the setup of the paradoxes (e.g. "this sentence is false" or "this sentence is unknown") do not *seem* to be meaningless. Furthermore, pointing to self-reference as the source of meaninglessness is also problematic as many self-referential sentences (e.g. "this sentence is in English") seem quite fine, and some of the semantic paradoxes can be formalized without self-reference.[2]

Even if it is not without difficulties, and even if it may not be quite fashionable nowadays, this is the approach I will pursue in this essay. The main objective of the essay is to introduce, and argue for, a Moderate Anti-Realist (MAR) framework, based on the verificationist/anti-realist principles that *neither truth, nor meaning can outstrip knowability*. I will start the first section with the Church-Fitch paradox that shows the limits of naïve (or radical) anti-realism. As a response to the paradox, I will introduce a MAR truth operator that defines truth – at least partially – in terms of *knowability*. Some of the relevant logical features of this truth-definition will also be discussed in the first section, among them how truths are different from mere facts (or factual propositions) and how this definition partitions propositions into eight classes, on the basis of their knowability. This division will motivate the question we will ask in the fourth section of the essay: given our anti-realist principles, what theory of propositional meaning can accommodate to our expectations?

The point of introducing a formal truth operator on the one hand, and a possible world interpretation of propositional content/meaning, on the other hand, is not to prove that anti-realism holds. Rather, the point is that such an approach is an adequate and efficient tool to solve a set of semantic paradoxes and other challenges.

I will briefly discuss the basic assumptions our MAR framework relies on in the second section. The individuation of propositions as the sets of possible worlds in which the given propositions are true or factual, as well as the shortcomings of this particular approach, will be discussed in the

[2] At least this is Yablo's claim (Yablo 1993), although it is not without its detractors.

third section. The fourth section will focus on the challenge of attributing meaning or meaninglessness to *fully unknowable* propositions, and the fifth section will offer a solution to this challenge. I will argue that the set of possible worlds that correspond to the meaning/content of propositions should contain only those worlds in which the proposition is not only factual but *known* as well.

I will apply this MAR framework to the *Liar Paradox* and the *Strengthened Liar* in the last section of the paper. I will demonstrate that our MAR framework can assign definite truth and factuality value to the Liar sentence, and that this truth/factuality assignment allows an explanation of the paradoxical nature of this sentence. I will suggest that the source of the paradox is that we try to attribute content/meaning to a sentence that is – given that it is unknowable – totally devoid of any meaning.

1. The knowability paradox and the MAR definition of truth

The generally agreed upon central tenets of antirealism are that neither truth, nor meaning can outstrip knowability. Somewhat more formally:

(VTP_{inf}) All truths are knowable, and

(VMP_{inf}) All meaningful propositions are knowable.

We will focus on the first of these principles in this section and return to the second one in the fourth section of this essay. The simplest, most straightforward way of formalizing VTP_{inf} is:

(VTP) $\vdash \forall p(p \rightarrow \Diamond Kp)$,

where the operator, K, should be read as "it is known that…".[3] The lesson of the Church-Fitch paradox (Fitch 1963), however, is that this straightforward formalization is inadequate. The paradox shows that, if – besides VTP – the factivity of knowledge and closure under conjunction-elimination in

[3] More formally, $K_{s,t}p$ is the operator that "the epistemic agent, s, knows that p at time, t." Then we can get the above K by generalizing over subjects (epistemic agents) and times: $Kp \leftrightarrow \exists s \exists t K_{s,t}p$

K are also granted,[4] then true propositions are not only knowable, but known as well:

(CFP) $\vdash \forall p(p \leftrightarrow Kp)$,

At the heart of the paradox is the following type of propositions:

(NC) $p \,\&\, \neg Kp$,

i.e. p is an unknown fact. While most of us would agree without hesitation that there are unknown facts, it is *impossible* to single out any of them and hence it is *unknowable* that a fact is unknown.

Another problem with VTP is that it provides only a *necessary*, but not a *sufficient* condition for truth, as the converse of VTP,

(VTP_{conv}) $\vdash \Diamond Kp \rightarrow p$,

does not hold, given that knowability ($\Diamond Kp$) is arguably not a factive.[5] Without a sufficient condition, however, there is a theoretical gap between knowability and truth – the epistemic-metaphysical element that would differentiate between these two concepts is missing.

One way to prevent the paradox, as I argued elsewhere (Marton 2006), is to revise the knowability principles (VTP and VTP_{conv}) in the following way:

---

[4] Formally: $\vdash \forall p(Kp \rightarrow p)$ for factivity, and $\vdash \forall p(K(p\&q) \rightarrow (Kp\&Kq))$ for conjunction-elimination. Then, here is how the paradox goes: assume, for any arbitrary p, that it is an unknown fact, i.e. $p \,\&\, \neg Kp$. If so, then it is knowable (by VTP), and so $\Diamond K(p \,\&\, \neg Kp)$. Then, in some possible world, $K(p\&\neg Kp)$. Given that knowledge is closed under conjunction-introduction, $Kp \,\&\, K\neg Kp$ also holds. So, given that K is a factive, a contradiction can be derived, and so we can discharge the assumption. Thus, it is $\neg(p \,\&\, \neg Kp)$, for any p, i.e. $\vdash \forall p(p \rightarrow Kp)$ holds. Given that K is factive, we can swiftly derive that $\vdash \forall p(p \leftrightarrow Kp)$.

[5] While the factivity of VTP_{conv} was accepted and argued for in the recent past, this principle now seems to be abandoned. See, e.g., Tennant's retraction, for the record (Tennant 2009, 225). Furthermore, even if one does accept the converse knowability principle, $\vdash \Diamond Kp \rightarrow p$, it leads to further paradoxes, as e.g. the modal collapse ($\vdash p \leftrightarrow \Diamond Kp$), if S4 is also granted (Williamson 1992).

(MART) $\vdash Tp \leftrightarrow (p\&\Diamond Kp)$,

where T is a *moderate anti-realist truth operator*. This definition introduces a distinction between truths and facts: while an unknown fact is indeed a fact, it is not a truth, as it is outside of our epistemic reach.

Truth, according to this definition, is essentially two-pronged: the second, epistemic part, $\Diamond Kp$, expresses its anti-realist ideals: truths are more than just facts out there; truths are essentially for us, epistemic agents. The first part, however, acknowledges that truths are not entirely within our realms – at the end, they are determined by how the world is. This, of course, is what "moderates" the anti-realist character of truth. Alternatively, truth is neither purely metaphysical/ontological, nor it is purely epistemic; these two aspects cannot be reduced to either one of them.

Introducing the MAR truth operator obviously preempts the Church-Fitch paradox as NC type sentences, i.e. sentences in the form: $p \& \neg Kp$, are not knowable, and so they are not true either. Our MAR interpretation of VTP_{inf} recognizes that this principle is about *truths*, and not facts in general.

In light of these insights, I will refer to propositions that hold in a given world as being *factual*, preserving the term "true" for propositions in the extension of our newly introduced operator, T. Obviously, all true propositions are factual, however not all factual propositions are true:[6] consider an *unknown contingent* statement, p; then, exactly one of the following two conjunctions must be factual as well: (i) $p\&\neg Kp$ or (ii) $\neg p\&\neg K\neg p$. But neither of them is knowable as they are NC-type propositions. Thus, some factuals are not true. The *logic of factuals* is the standard 2-valued classical logic where e.g. $p \vee \neg p$ is a theorem.

The *logic of truths*, however, is different. First, we can introduce the concept of falsity, mirroring the definition of truth, as follows:

(Def-F) $\vdash Fp \leftrightarrow (\neg p \,\&\, \Diamond K\neg p)$.

We can also notice that

---

[6] In other words, MART restricts *capture* while accepts *release* without any further ado.

$\vdash Fp \leftrightarrow T\neg p$,

as it can be expected. Given that certain propositions are neither true nor false, this system is a 3-valued logic embedded in the more general, bivalent system of factuals.

We can even go one step further: the concepts of truth and falsity were constructed from 3 logically independent elements: a proposition, p; its knowability, $\Diamond Kp$; and the knowability of its negation, $\Diamond K\neg p$. From these three ingredients we can manufacture eight different *classes* of propositions:[7]

(i) two classes for true propositions, i.e. propositions that are factual and their factuality is knowable:
- propositions that satisfy $p \,\&\, \Diamond Kp \,\&\, \Diamond K\neg p$, or
- propositions that satisfy $p \,\&\, \Diamond Kp \,\&\, \neg\Diamond K\neg p$.

(ii) two for false propositions, i.e. propositions that are non-factual and their non-factuality is knowable:
- propositions that satisfy $\neg p \,\&\, \Diamond Kp \,\&\, \Diamond K\neg p$, or
- propositions that satisfy $\neg p \,\&\, \neg\Diamond Kp \,\&\, \Diamond K\neg p$.

(iii) the remaining four for the 3rd value propositions, i.e. propositions whose (non-) factuality is unknowable:
- propositions that satisfy $p \,\&\, \neg\Diamond Kp \,\&\, \Diamond K\neg p$, or
- propositions that satisfy $p \,\&\, \neg\Diamond Kp \,\&\, \neg\Diamond K\neg p$, or
- propositions that satisfy $\neg p \,\&\, \Diamond Kp \,\&\, \neg\Diamond K\neg p$, or
- propositions that satisfy $\neg p \,\&\, \neg\Diamond Kp \,\&\, \neg\Diamond K\neg p$.[8]

---

[7] Belnap (1977, 47) considers a structurally similar eightfold division of propositions that also combines epistemic and ontological aspects in a similar way.

[8] There is another way to group these eight basic types:

(i) two of them are not only contingent, but *epistemically contingent*, i.e. both p and ¬p are knowable (propositions that satisfy $p \,\&\Diamond\, Kp \,\&\, \Diamond K\neg p$ and $\neg p \,\&\, \Diamond Kp \,\&\, \Diamond K\neg p$).

(ii) Two of them are *epistemically undisputable* i.e. they are true (or false) but their negation cannot be known (($p \,\&\, \Diamond Kp \,\&\, \neg\Diamond K\neg p$ and $\neg p \,\&\, \neg\Diamond Kp \,\&\, \Diamond K\neg p$) – necessary statements definitely do belong to this category, but arguably there are other propositions in this category as well.

We will soon ask: What kind of theory of propositional meaning can adequately render *meaninglessness* to *fully unknowable* propositions, i.e. to propositions that satisfy p & ¬◊Kp & ¬◊K¬p or ¬p & ¬◊Kp & ¬◊K¬p?

2. Basic Assumptions

We have to pause at this point in our investigation to address the basic underlying assumptions of our approach. First, we assume that only standard possible worlds (i.e. worlds without contradictions or value gaps) are in the set of all possible worlds. We will follow Kripke's approach (Kripke 1980), according to which possible worlds are not discovered, but rather stipulated.

Second, I will take S5 to be the relevant modal system (containing exactly one equivalence class). This choice gives us the comfort of equating possibility with being true/factual in at least one possible world without further specifying the accessibility relation.

Third, the relevant modality to be considered here is *logical* possibility and necessity. This choice of modality is forced upon us by our inquiries into the concept of meaning in the next sections of the essay; to properly represent the content of propositions, all possible worlds (not only those within some narrower concepts of modality such as nomological or metaphysical) must be considered. Arguably, however, the scope of modality is effectively narrowed by our use of the knowledge operator, K, as this operator is limited to (our kind of) epistemic agents.

Fourth, our MAR definition of truth requires a robust concept of knowledge. Unless this theoretical concept of knowledge strongly overlaps with our pre-theoretical, practical concept of knowledge, the truth definition has little use. This concept of knowledge should cover empirical, as well as theoretical knowledge, etc. I also take it for granted that an agent's knowing a proposition, p, implies its factuality (and so, its truth), and that

(iii) Two of these types are *epistemically disputable* or *falsifiable* (p & ¬◊Kp & ◊K¬p and ¬p & ◊Kp & ¬◊K¬p); while their factuality cannot be known, if they were non-factual, then their non-factuality could be known (p & ¬◊Kp & ◊K¬p and ¬p& ◊Kp & ¬◊K¬p).

(iv) Finally, two of these classes are *fully unknowable* (p & ¬◊Kp & ¬◊K¬p and ¬p & ¬◊Kp & ¬◊K¬p).

p is believed by the agent. A third condition involving some kind of justification, reason, or evidence is also assumed. It may be objected that this assumption is overly optimistic as we have no generally accepted, adequate *theory* of knowledge. But this criticism conflates two different issues; namely, the lack of a theory for a concept with the viability of the concept itself.

Fifth, this essay will avoid treating the concept of knowledge as an essentially modal concept, and accordingly, the knowledge operator, K, as a modal operator. In other words, this essay will not utilize the nowadays popular, formal 2-dimensional approaches, where knowing p in a given world is essentially a function of whether or not p is true in the epistemologically accessible worlds. These formal models, no doubt, have their relevance in certain epistemological investigations. But those models also come with their own limitations and problems (e.g. that any necessary proposition is known, according to the modal interpretation of the knowledge operator). It is also rather doubtful whether these models are consistent with the basic ideals of anti-realism.

Finally, knowledge claims (i.e. that s knows that p at t, $K_{s,t}p$, and the more generalized form, it is known that p, Kp) are epistemic facts, and as such they are parts, or constituents of possible worlds, and can be expressed by propositions. In other words, epistemic facts are facts, and the corresponding propositions can be individuated the same way as any other propositions, i.e. by the corresponding set of possible worlds.

3. On the meaning/content of propositions

It is generally accepted in certain philosophical circles that propositions can be individuated and differentiated by the sets of possible worlds in which the corresponding propositions are *factual*. If two propositions, p_1 and p_2, have different truth (or rather, factuality) values in at least one possible world, then the two propositions are indeed different. However, if there is no such world, then p_1 and p_2 are just two instances of the *same* proposition.

By identifying propositions with sets of worlds, the *meaning* of these propositions is intended to be captured. Indeed, one way to understand propositions – which amounts to capturing their meanings – is to ask: in

what circumstances is this proposition true/factual[9] and in what circumstances is it false/non-factual?

This way of identifying the meaning of propositions is not without difficulties. First, there is the threat of circularity: we define the meaning of a particular proposition by referencing some relevant situations which, one can presume, are identified by some other propositions. But those situation-describing propositions (or, truth and falsity conditions) must be individuated and interpreted in some way[10] and that seems impossible to do without referencing – sooner or later – the particular propositions we have started with. However, even if defining meaning this way is circular, it does not necessarily mean that it is *viciously* so.[11]

Second, differentiating propositions by sets of possible worlds results in the fact that there is exactly one necessary proposition. Still, even if a formal individuation leads to the outcome that there should be only one necessary proposition, there should be some differentiation among its instances, according to their differing meanings. To wit, the proposition that "if you don't stand for anything, then you don't stand for anything" means something entirely different than the proposition that "two plus two equals four." This problem, the problem of hyperintensionality[12] is outside of the scope of this essay – but it definitely motivates the position about the meaning of propositions I will argue for.

Finally, as I have indicated earlier, the meaning/content of fully unknowable proposition is also problematic. What is the relevance of differentiating two propositions if both are unknowable to us? Alternatively, our

---

[9] Given that our preferred modal system is S5, the difference between being true and being factual plays any role only if the proposition is *fully unknowable*. If a proposition is known in at least one possible world, then the proposition is true in all those possible worlds in which it is factual. As the issue of fully unknowable propositions will be considered soon in some length, the distinction will be downplayed here.

[10] This is the result of our dependence on Kripke's take on the ontological status of possible worlds – or at least of the way I interpret his claim.

[11] To substantiate this point, a Quinean argument for the primacy of theory over the individual sentences may be handy here. And surely, advocates of coherence theories of truth and/or of knowledge (e.g. Davidson 1986), can also help here. However, this issue has little significance for our project and so it will not be pursued on these pages.

[12] On this problem, see e.g. Jago (2014).

MAR approach is built – at least partially – on the idea that meaning should not outstrip knowability. In light of this consideration, we may ask: what kind of theory of meaning can accommodate to this anti-realist expectation?

4. On the meaning of unknowable propositions

Let us consider a fully unknowable, contingent proposition, p. This proposition, like any other, can be individuated by the set of possible worlds in which p is factual. But what content/meaning does p have? As mentioned earlier, identifying meaning with the set of possible worlds appeals to our intuition that the meaning of propositions can be grasped by the situations in which they are true, and the situations in which they are false. Ideally, we could assemble a list of propositions, p_1, …, p_n such that $(p_1 \& p_2 \& \dots p_n) \leftrightarrow p$. This list of propositions, p's truth conditions, would then explicate the meaning of p. Given that this biconditional, $(p_1 \& p_2 \& \dots p_n) \leftrightarrow p$, fixes the *meaning* of p, it should hold not only in the actual world, but in every possible world as well.

One may find it more natural to identify a given proposition not with the *conjunction*, but rather the *disjunction* of a set of propositions. As I see it, both options are viable, but they are motivated by very different considerations. The latter option envisions the identification of a proposition, p, with listing all the possible scenarios in which p is factual. Accordingly, each disjunct is a detailed description of a possible world (or perhaps a narrowly defined situation, i.e. a "small" set of possible worlds). The former option, more fitting for an anti-realist approach, looks for defining criteria, such that each criterion is necessary, and they are jointly sufficient as well. Actually, we have already utilized this approach earlier, when we defined or identified the concept, or rather the *meaning*, of truth with two individually necessary and jointly sufficient conditions: factuality and knowability. Similarly, the traditional justified true belief approach to knowledge that we listed among our assumptions in the previous section also follows this pattern.[13]

[13] Propositions can, and perhaps should, be identified by sets of propositions. There are two different strategies to identify sets: either by listing their elements, or by giving

In practice, we would probably settle for less: we would describe a scenario (i.e. a collection of worlds) in which p is true and another scenario (i.e. another collection of worlds) in which p is false. These scenarios can be described by sets of propositions, $q_1, q_2, \ldots, q_l$ and $r_1, r_2, \ldots, r_m$ and then we would claim that $(q_1 \,\&\, q_2 \,\&\, \ldots \,\&\, q_l) \rightarrow p$ and $(r_1 \,\&\, r_2 \,\&\, \ldots \,\&\, r_m) \rightarrow \neg p$. These two sets represent the truth and falsity conditions of p – hereafter the T&F conditions. Intuitively, the problem is that if p and ¬p are unknowable, then so are the T&F conditions that meant to explicate them. But how can we grasp the meaning of a proposition if it is couched in descriptions that are themselves unknowable? Of course, the meaning of the propositions constituting the lists can be further explained by further sets of lists, but the same must hold true: some of those propositions on those lists must also be unknowable, and so on.[14]

To support our intuition, I will briefly argue first that if the T&F conditions are known, then p cannot be unknowable. Then we will consider in what ways these conditions themselves can be unknowable. The two conditionals, presenting a conceptual analysis for the meaning of p, are *analytic* as they explicate meanings and they are obviously *known* in our world. There are two ways these conditionals can fail to transfer knowledge from T&F conditions to p:

(i) in all the possible worlds where the world-describing T&F conditions are actually known, the conditionals themselves are not known, or

(ii) although the conditionals themselves are known, but epistemic closure does not hold. Neither of these options are reasonable, though.

---

rules. The former corresponds to the *disjunctive* approach, while the latter to our preferred approach of using *conjunctions*.

[14] Let me acknowledge two possible, highly connected, objections – at least in a footnote – which will not be discussed here. First, one may object that there are no unknowable facts, i.e. all facts are within our epistemic reach. Second, that even if there are unknowable facts, there are no unknowable propositions to express them. As I do not think that these objections are reasonable, they will not be discussed here.

Considering the former option, we may ask: is there any reason to assume that in all the possible worlds where the T&F conditions, (q_1, q_2, …, q_l) and (r_1, r_2, …, r_m), are fully known, the analytic conditionals, (q_1 & q_2 & … & q_l) → p and (r_1 & r_2 & … & r_m) → ¬p, known in our own actual world, would not, or rather could not, be known (at any time, by any epistemic agent)? Since these conditionals are true in any world, it is either the belief or the justification/evidence condition that can prevent putative knowers from knowing them. If so, then there must be some *inherent, structural* difference between our world in which these knowledge conditions are met and the worlds in which the T&F conditions are known to account for the difference in knowing the analytic conditionals. But, as far as I can see, there is no such inherent difference.

Turning our attention now to the latter option, these conditionals can fail to transmit knowability only if epistemic closure itself is challenged: these challenges operate with a familiar line of reasoning, summarized in a conditional, against a non-standard, unexpected circumstance (Dretske's zebra-looking mules, etc.). But our conceptual analysis does not fit into this pattern – it outlines the very circumstance in which the analyzed concept must be true. If both the T&F conditions are knowable and the analytic conditionals are known, then transmitting knowability is unavoidable.[15]

Accordingly, if p is unknowable, then the set of T&F conditions must also be unknowable. There are 3 different ways these T&F conditions can be unknowable:

(i) The set of T&F conditions is inconsistent, and so the conditionals are vacuously true and p and ¬p may be propositions out of our epistemic reach.

(ii) Even if the T&F set is consistent, some of the elements of the sets can be unknowable themselves, and that accounts for the unknowability of p.

---

[15] This claim may come with a caveat. It may be objected that even if both K(q_1 & q_2 … & q_l) and K((q_1 & q_2 … & q_l) → p) hold, epistemic agents may never actually attain the knowledge of q. That's certainly *possible*, i.e. there will be possible worlds in which p will not be actually known. But if closure holds, then attaining the knowledge of p is also *possible*, i.e. there will be possible worlds in which p is known. And that is enough for our purposes.

(iii) Even if the set of T&F conditions is consistent, and all the propositions in this set are knowable *in themselves*, their conjunctions may still not be knowable, i.e. the situation they describe are not fully knowable. NC, the sentence at the heart of the Church-Fitch hypothesis is an example of such conjunctions where both sentences can be knowable separately, but the conjunction itself is arguably unknowable. Importantly, this option undermines the compositionality of meaning – even if two propositions, p and q, both have meanings, p&q may not be knowable and thus the conjunction is meaningless.[16]

To sum it up, unknowable propositions cannot be explicated/illuminated/interpreted in terms of knowable propositions. They are meaningless, as they are beyond our epistemic reach. As these propositions may be individuated by a corresponding set of possible worlds, we can further conclude that meanings cannot be identified with the sets of possible worlds in which the proposition is factual, even in case of contingent propositions.

5. A solution to the problem

The insights of the previous section suggest that we should refine our intuition about the individuation and meaning of propositions by modifying our previous question in the following way: in what *knowable* circumstances would a proposition be *known* to be true, and in what *knowable* circumstances would a proposition be *known* to be false? Accordingly, we

[16] Jago writes: "Take our example from above, 'it is both snowing and not snowing here right now'. This sentence is perfectly meaningful, for both of its conjuncts are meaningful, and a sentence 'A ∧ B' is meaningful whenever its conjuncts 'A' and 'B' are individually meaningful" (Jago 2014, 7). Jago's claim about the compositionality of meaning comes without any argument or support. However, just because one understands the meaning of the proposition "it's snowing here right now," claiming that the proposition that "it is both snowing and not snowing here right now" has any meaning is far from obvious. Personally, I cannot imagine what would anyone aimed to express by that proposition. Furthermore, it is unclear how the meaning of this proposition is different from "Boston is in Massachusetts, but Boston is not in Massachusetts."

identify the *meaning* of a proposition not with the set of worlds in which the proposition is factual, but rather with the set of worlds in which the proposition is *known*. Informally, this approach emphasizes the relevance of context – the meaning of a proposition, p, is captured by considering *what else should be known* to understand p.[17]

In essence, our MAR approach suggests *two different identity relations* on the set of propositions. In one way, propositions can be individuated by the sets of worlds in which they are factual. In another way, the content/meaning of propositions can be identified with set of possible worlds in which the proposition is known. These two different identity relations correspond to our two different concepts of truth; the metaphysical concept (our concept of factuality) and the epistemically constrained MAR concept (our concept of truth).

To be more precise, it is not propositions, but rather pairs of propositions to which meaning is attributed. In the traditional account, individuating p with the corresponding set of worlds also individuates ¬p, as its corresponding set is the complement set. Accordingly, the meaning of the pair of propositions, p and ¬p (or rather, Tp and Fp) should be identified with the set of worlds in which p is known and with the set of worlds in which ¬p is known. Quite obviously, this approach solves our problem. If p and ¬p are both unknown in every possible world (i.e. p is a fully unknowable proposition), then the corresponding sets are empty and so no meaning is associated with p (and ¬p). It also explains in what sense truth is more than mere factuality – being true is being meaningfully factual.

6. The Liar Paradox

Let us turn our attention now to the Liar Paradox. Consider first the following sentence:

---

[17] This point suggests how this approach can solve the problem of hyperintensionality. What defines the meaning of a necessary statement is the set of worlds in which that statement is known. Focusing on those worlds would reveal what should have to be known previously to be able to know that p. This approach is rather similar to the intuitionistic ideal of stages, or possible development, of knowledge (Beall 2003, 96-97).

(f) This sentence is false.

Traditionally, this sentence can be formalized as

(1) $f \leftrightarrow \neg f$,

and then it is easy to realize that no truth value can be attributed to f. This point, on its own, invites us to consider some version of a 3-valued (or many-valued) logic. The trap can be easily avoided if a 3rd value is attributed to both f and ¬f. Still, some explanation is required about the meaning of the 3rd truth value, i.e. what it means for a proposition to be neither true nor false.

Consider now the following, "strengthened" version of the paradox:

(n) This sentence is not true.

It is often claimed that the previous approach, based on a 3rd truth value, is inefficient here. If n is neither true, nor false, then obviously n is not true, so the sentence that "n is not true" *is* true and so we are back at the paradox. We may anticipate at this point that this conclusion is just too fast; all we should be able to conclude from this reasoning is that "n is not true" *is factual*.

Let us now switch from the traditional approach to our MAR approach. The previously introduced sentence, f, can be written as

(2) $f \leftrightarrow Ff$,

Where F is our *falsity* operator. Since (2) fixes the meaning of the sentence referenced as f, it is an analytic statement and so

(3) $\Box(f \leftrightarrow Ff)$.

Given that Ff is defined as $\neg f \,\&\, \Diamond K\neg f$, what (3) really amounts to is

(4) $\Box(f \leftrightarrow (\neg f \,\&\, \Diamond K\neg f))$.

As f and $\neg f \& \Diamond K\neg f$ cannot be simultaneously factual, (4) can only be true if both are nonfactual. Accordingly, f must be non-factual, and so ¬f must

be factual, but then ◊K¬f cannot be factual. Putting all these considerations together, 4 implies that

(5) □(¬f & ¬◊K¬f).

Furthermore, according to 5, ¬f is factual in all possible worlds and thus f is not knowable either:

(6) □(¬f & ¬◊Kf & ¬◊K¬f).

Less formally, the crucial sentence in the liar paradox is, on the one hand, nonfactual, but, on the other hand, it is also *fully unknowable*, and so *meaningless*.[18]

Similarly, the formal version of the "strengthened" liar sentence is

(7) □(n ↔ ¬Tn),

and then, by using the definition of T:

(8) □(n ↔ ¬(n & ◊Kn)).

Following a reasoning similar to our previous one shows that (8) implies that

(9) □(n & ¬◊Kn & ¬◊K¬n).

(9) shows that the crucial sentence of the *strengthened liar* is factual, but, just like the *liar*, it is also fully unknowable,[19] and so meaningless.

---

[18] Even if we know f's (non-) factuality (that it holds in none of the possible worlds), f itself is unknown to us. This point brings to the front one of the basic tenets of antirealism: knowledge of meaning is knowledge of truth conditions. Knowing the factivity of a proposition is not the same as knowing its meaning, or, simply put, knowing the proposition itself.

[19] Here is the reason: assume that the sentence within the scope of the necessity operator, n & ¬◊Kn & ¬◊K¬n, is knowable. Then there is a world, w, in which K(n & ¬◊Kn & ¬◊K¬n) is factual/true. Assuming that K is closed under conjunction-elimination, both Kn and K¬◊Kn would hold in that world. Given that K is factive, ¬◊Kn (i.e. □¬Kn)

Our outcomes, (6) and (9), show that the MAR definitions of truth and falsity render definite factuality and truth values to the *liar* and *strengthened liar* sentences. The former is unknowably nonfactual, while the latter is unknowably factual; as such, neither of them is either true or false.

Our language is limited by the logical/epistemic norm that sentences should be meaningful. Any assertion in the form that "p is factual" should be interpreted as p is *meant to be meaningfully factual*, which, according to our interpretation means that p is meant to be true. If so, then the distinctions between untrue and false and, similarly, between true and unfalse disappear when we consider assertions. False and untrue sentences differ in their knowability and no one should aim to assert a proposition that's unknowable, and thus, according to our reasoning, meaningless. The difference between these sentences surfaces only as an explanatory device: they explain one's mistake to assert something that should not or could not be asserted. Arguably, this is an important advantage of this approach: the distinction between truth and factuality is almost imperceptible – in most of our everyday (and perhaps even in our theoretical) dealings truth and factuality are the same. The consequence of this insight is that the two liar paradoxes collapse into one; if there is no real, *meaningful* difference between the assertions that "this sentence is false" and that "this sentence is untrue" then these assertions express the same proposition that can be expressed, using the combination of our previous formulations, (6) and (9), as follows:

(10) $\Box((f \,\&\, \neg\Diamond Kf \,\&\, \neg\Diamond K\neg f) \vee (\neg f \,\&\, \neg\Diamond Kf \,\&\, \neg\Diamond K\neg f))$.

In other words, the proposition that expresses the liar-sentence is the staple meaningless sentence which is either untruly factual *or* unfalsely nonfactual.

Furthermore, our insight about the normative character of assertions prevents the emergence of an "iterated" liar paradox. In light of our previous insights about the *normative* standards governing assertions, the sentence that

---

would also hold in that world, and so would ¬Kn then. Given that both Kn and ¬Kn would both hold in w, n & ¬◊Kn & ¬◊K¬n is not knowable.

(g) this sentence is nonfactual,

should be interpreted as "this sentence is meaningfully nonfactual", i.e. it is false. But then we are back at the original liar paradox.

7. Concluding remarks

The underlying, fundamental assumption of this paper is that the concept of knowledge plays a central role in our concepts of truth and meaning. It is the possibility of knowing that makes truth (understood as being different than mere factuality) and meaning possible for us, epistemic agents. I did not argue for this fundamental assumption in this essay. Rather, I argued that such a moderate anti-realist approach to the concepts of truth and meaning offers a way of solving or dissolving a number of semantic paradoxes and other challenges. I demonstrated that our MAR approach to truth and meaning, offers an interpretation of the *Liar* and *Strengthened Liar* sentences. This interpretation renders *definite* truth and factuality values to these sentences and shows that the difference between their truth and factuality values emerge from their *meaninglessness*. Even if these sentences do not seem to be meaningless, they are meaningless as there is no possible situation in which these sentences (or their negations) could be known. This should not surprise us: if someone tells us that the sentence she is uttering is false (or it is not true), then we would not know what she meant by it, what kind of knowable fact she tried to impart to us.

References

BEALL, J. C. & VAN FRAASSEN, B. C. (2003): *Possibilities and Paradox*. Oxford: Oxford University Press.

BELNAP, N. D. (1976): How a Computer Should Think. In: Ryle, G. (ed.): *Contemporary Aspects of Philosophy*. Stocksfield: Oriel Press, 30-56.

DAVIDSON, D. (1986): A Coherence Theory of Truth and Knowledge. In: LePore, E. (ed.): *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Oxford: Basil Blackwell, 307-319.

FITCH, F. B. (1963): A Logical Analysis of Some Value Concepts. *The Journal of Philosophical Logic* 28, 135-142.

JAGO, M. (2014): *The Impossible. An Essay on Hyperintensionality.* Oxford: Oxford University Press.

KRIPKE, S. (1980): *Naming and Necessity*. Harvard University Press.

MARTON, P. (2006): Verificationists versus Realists: The Battle over Knowability. *Synthese*, 151, 81-98.

TENNANT, N. (2009): Revamping the Restriction Strategy. In: Salerno, J. (ed.): *New Essays on the Knowability Paradox*. Oxford: Oxford University Press.

WILLIAMSON, T. (1992): On Intuitionistic Modal Epistemic Logic. *Journal of Philosophical Logic* 21, 63-89.

YABLO, S. (1993): Paradox Without Self-Reference. *Analysis* 53, No. 4, 251-252.

Organon F 25 (2) 2018: 174-195

Inferentialism without Normativity

KRZYSZTOF POSŁAJKO – PAWEŁ GRABARCZYK[1]

ABSTRACT: In this paper we argue that inferentialist approach to meaning does not, by itself, show that meaning is normative in a prescriptive sense, and that the constitutive rules argument is especially troubling for this position. To show that, we present the proto-inferentialist theory developed by Ajdukiewicz and claim that despite the differences between his theory and contemporary inferentialism rules of language in both theories function more like classificatory devices than prescriptions. Inferentialists can respond by claiming that in their theory meaning is essentially social and hence normative, but we claim that then semantic normativity becomes derivative of social normativity.

KEYWORDS: Ajdukiewicz – Brandom – inferentialism – normativity of meaning.

[1] Received: 8 January 2018 / Accepted: 12 March 2018

✉ Krzysztof Posłajko

Department of Philosophy, Jagiellonian University
Grodzka 52
31-044, Krakow, Poland

e-mail: krzysztof.poslajko@uj.edu.pl

✉ Paweł Grabarczyk

University of Lodz
Narutowicza 68 street
90-136, Lodz, Poland

e-mail: pagrab@gmail.com

1. Introduction

The aim of this paper is to show that the argument from constitutive rules provides a substantial challenge for the idea that inferentialist theory of meaning implies prescriptive normativity of meaning. The constitutive rules argument was formulated as a general objection to the thesis that there is a prescriptive normativity involved in meaning ascriptions. Our goal is to show that this argument is especially problematic for the adherents of inferentialist account of meaning, who usually subscribe to the normativist position in the normativity of meaning debate. In order to show that, we will present Ajdukiewicz's theory of meaning, which, as we believe, provides a useful, albeit slightly simplified model of how defining meaning in terms of inferences leads to the conclusion that the putative semantic norms are constitutive (and henceforth non-prescriptive). Finally, we present a way in which an inferentialist may refute this argument and claim that it works only if we assume that normativity of meaning is derivative of social normativity.

2. Inferentialism and normativity of meaning

Inferentialism is a variant of the broadly conceived inferential/conceptual role semantics. In the most general sense, inferential/conceptual role semantics is a doctrine saying that the meaning of an expression depends on the function this expression has in inferences (Whiting 2015). Thus understood, inferential role semantics is a subspecies of functional role semantics, which connects the notion of meaning of a term with the function this term plays in a language.

Inferentialism, when properly conceived, has its roots in the theory of Sellars (see his 1954, 1973, 1974). Currently, perhaps the most influential proponent of inferentialism is Robert Brandom (1994, 2000), but several other thinkers espouse some sort of affinity towards this theory – among them there are Michael Williams (2013), Matthew Chrisman (2010), Alexis Burgess (2015), and Jaroslav Peregrin (2012, 2014).

The claim that meaning is normative is often seen as an essential feature of inferentialism. This claim is made by both: proponents of inferentialism (Brandom 1994, Peregrin 2012, Shapiro 2004) and critics of this approach

(see e.g. Hattiangadi 2003). According to Brandom "the propositional contents (…) are conferred on expressions, performances, attitudes, and statuses by their playing a suitable role in a system of discursive normative social practices" (Brandom 1994, 63-64).

Recently, Peregrin (2014, 8-9) has stated that the claim of normativity of meaning is a defining feature of inferentialism, as accepting the normative character of rules distinguishes inferentialism from inferential role semantics. According to the causal version of inferential role semantics (see Boghossian 1993), what determines meaning is a network of actual dispositions for making inferences that users of a symbol possess. Inferentialists, on the other hand, tend to see the inferences which define the meaning of an expression as inferences which are *correct* and which the users *should* make.

This is closely related to the second aspect that differentiates inferentialism from inferential role semantics, namely the inferentialists' claim that meanings are constituted on a social level. These two aspects of inferentialism – treating linguistic rules as normative and meanings as social – are logically independent but they seem to go together quite naturally: if one is keen to claim that meaning is a set of correct inferences, then one would be tempted to say that the correctness is somehow determined by societal standards.

It might then seem that we have a clear distinction between two kinds of inferential-based semantics – one is individualistic and dispositional, hence descriptive, and the other – inferentialism – social and normative. However, this clear picture could be undermined if it were possible to show that meaning defined on the grounds of the inferentialist theory does not have to be a normative notion. In what follows, we should try to construe an argument to the effect that – despite what inferentialists officially proclaim – it is not easy to claim that meaning does not have to be normative in their theoretical framework, at least in the sense of "normativity" that many inferentialists have assumed.

The question whether meaning is normative has been a subject of ongoing controversy during the last three decades (see Glüer & Wikforss 2016 for an overview). Initially, the normativist position gained widespread acceptance (Kripke 1982, McDowell 1984, Boghossian 1989). Later, the anti-normativist position started to undermine the initial normativist consensus (Glüer & Pagin 1998, Wikforss 2001, Hattiangadi 2007),

although many philosophers still defend normativism (Glock 2005, Whiting 2007).

The normativist stance was initially based on the observation that people normally asses their own and other people's utterances as correct or incorrect (see e.g. Kripke 1982). This fact is usually taken to be uncontroversial and nearly all participants in the normativity of meaning debate (with the possible exception of Davidson 2005) seem to accept the premise that there is a sense in which utterances can be characterized as semantically correct or not.

What is contested by the anti-normativists, however, is the fact that this correctness amounts to "genuine normativity". This is, for example, clarified as a claim that the normativity involved in meaning is not "genuinely prescriptive" – we can say that a certain way of speaking is "semantically correct" but this does not provide anyone with any reason to act with accordance to the relevant semantic rule (see e.g. Hattiangadi 2007).

The difference between prescriptive and other kind of rules has been widely acknowledged in the literature on normativity of meaning (see e.g. Whiting 2007). It is uncontroversial that it is admissible to characterize certain linguistic behaviour as correct or not. However it is usually claimed that there is a logical difference between the claim that there are certain norms which allow us to say that something is correct or not, and the logically stronger claim that one ought to act the correct way. Let us consider a trivial example: there might be norms that state that an infant of six months of age has a "correct" weight only if it weights between 6 and 11 kilograms. Still, one would not say that a child should weigh between 6 and 11, in the sense that there is an obligation of any sort for the child to have the appropriate weight.

Normativists sometimes respond with the observation that the way we use normative vocabulary in linguistic context is perfectly valid from the folk point of view (see e.g. Glock 2005). In response to that, anti-normativists claim that meaning is not normative in the prescriptive sense of the term. They also claim that the fact that it might be described in normative vocabulary, is not, in itself, a "philosophically interesting" thesis. Boghossian (2005), for example, stresses the point that the normativity of meaning claim might be true on certain reading, but this reading makes the claim trivial; if the claim were to be interesting it must show that meaning is normative in genuinely prescriptive sense. Anti-normativists might claim

that the sense in which meaning can be normative is of no use when we consider the question of naturalization of meaning (this seems to be the guiding idea of Hattiangadi's 2007). In a similar vein, Miller (2010, 2012) argues that classical arguments for ethical anti-realism do not apply in the semantic case, as the normativity in the semantic case is distinct from the ethical one.

The proponents of the strong version of normativity of meaning thesis state that not only can we claim that some uses of certain expressions are correct or not but that this observation warrants the claim that users ought to use these expressions in a way that is semantically correct (Whiting 2007). So, according to the strong normativist stance, there is a straightforward connection between the fact that meaning comes along with correctness conditions and the claim that meaning entails prescription.

Anti-normativists deny this. A crucial element of the anti-normativist strategy is to explain this "non-interesting" sense of "correctness" which can be applied to semantic claims. There are several ways in which this can be done. In what follows, we are going to focus on a strategy which uses the notion of constitutive rules, because we believe it is the most potent one in the context of inferentialist approach to meaning.

3. Constitutive rules challenge

The constitutive rules strategy, developed by, among others, Glüer and Pagin (1998) and Wikforss (2001), amounts to the claim that statements that express the putative norms of semantic correctness are not genuinely prescriptive rules, as they are just constitutive rules. The notion of constitutive rule has been popularized by Searle (1969), according to whom we should distinguish between two basic kinds of rules. Prescriptive rules regulate already preexisting behaviours, whilst constitutive rules constitute new ones, in the sense that certain physical actions become classified as some institutional ones. Constitutive rules are those which define what kinds of behaviours count as kinds of doings in certain contexts: for Searle the canonical form of constitutive rule is "Action A counts as doing B in context C". This means that a rule of constitutive kind is used to say which actions one should undertake, if one wants to perform certain institutionally or socially defined deed. A primary example here are the rules of chess,

especially such rules which define what kind of move count as, say, castling.

Constitutive rules are not prescriptive in any meaningful sense. The rules of chess do not dictate which moves one should make in certain situations nor that one should play chess at all (one might easily use the same pieces to play an entirely different game or just throw them idly). What constitutive rules do is that they specify what kinds of doings would count as playing a game of chess and making specific moves in that game. Needless to say, there are many rules which prescribe the right moves in the right situations. Apart from constitutive rules, there are rules which teach the players to play chess well (as opposed to teaching them how to play chess at all).

Should semantic rules be indeed constitutive, it would mean that they do not provide any prescriptions concerning the use of words. Rather, semantic rules should be taken to constitute the meanings. As Wikforss puts it:

> According to this picture, there is a constitutive relation between use and meaning such that in order to mean horse by "horse" you must use (be disposed to use) your words in certain ways. The 'must' here, again, is not an 'ought' in disguise; it is not the 'must' of a prescription. (Wikforss 2001, 218)

One might wonder, however, why there is a conflict between semantic rules being prescriptive and semantic rules being constitutive. Glüer and Pagin answer this question by pointing out that constitutive rules do not explain action in a relevant way. A constitutive rule, as standardly conceived, "does not occupy a motivational position in the practical argument. It occupies a doxastic position, that is, it functions just as an ordinary belief in effecting a theoretical transition from one pro-attitude to another" (Glüer & Pagin 1998, 218).

According to a standard philosophical story, an intentional explanation of action (the so-called practical syllogism) necessarily involves two "premises" – one which is motivational (a desire to achieve X) and the other which is factual/descriptive (action A will result in achieving X). According to the Humean story of the normative (see e.g. Smith 1994), normative statements enter the reason-based explanations in the motivational

position. This is one of the main sources of the problem with treating normative statements as descriptive.

Why constitutive rules cannot enter the motivational role? The answer Glüer and Pagin provide is quite convincing. If the content of a constitutive rule is given by the formula "Action A counts as doing B in context C", this is by no means motivating for anyone to do A. What is needed is an additional motivational premise that one should aim for doing B in context C – this motivational premise can be a simple desire or it can be some normative premise stating that one should aim to do B in context C (for, say, moral reasons).

The contrast between constitutive and prescriptive rules might be illustrated by looking at social norms. There are certainly social norms which are constitutive in character, like the norms which specify what kind of things needs to be done in order to marry (like signing an appropriate documents). Other social norms might be prescriptive, like the norms of etiquette, which might state that the wedding couple should dress formally. The difference between the two kinds of rules is best seen if we look at what happens when they are violated. If one does not sign the appropriate documents then there is no marriage ceremony; however, if the couple attends their own wedding in old Nirvana t-shirts, this does not invalidate the marriage (although it might be deemed inappropriate).

Games also contain "constitutive rules". Instead of differentiating between valuable and invaluable moves, they help us decide if actions are to be classified as belonging to the game. Apart from good and bad moves in chess, there are also invalid moves. Even though from the physical point of view the player can make illegal moves, she cannot, as it were, make them in the game, because they will be instantly classified as not belonging to the game.

Morality, on the other hand, is usually thought to consist of prescriptive rules. Moreover, it is also claimed that moral rules are "objectively prescriptive" – in the sense that they provide prescriptions which are independent of any contextual factors and individual desires (see e.g. Boghossian 2005). If morality is objectively prescriptive, whilst semantic norms are constitutive, then it might be said that the normativity of linguistic norms is different from the normativity enjoyed by the moral.

Thus, the constitutive rules argument may be summarized as follows:

First premise:	semantic rules are constitutive rules;
Second premise:	constitutive rules are not genuine prescriptive rules;
Conclusion:	semantic rules are not genuine prescriptive rules.

Again, it is important to note that this argument does not aim to show that semantic rules are not normative in any sense. Rather it shows that it is not normative in the technical sense, assumed in many debates in contemporary metaethics.

Inferentialists seem to have a tendency to downplay the importance of the argument from constitutive rules. The opinion voiced by Peregrin seems to be characteristic of this approach: "The fact that the rules constitute meanings does not rob them of their normativity" (Peregrin 2012, 96). The offshoot of Peregrin's discussion on the constitutive rule argument seems to be that for many inferentialists there is nothing inconsistent in the thesis that semantic rules can be constitutive and genuinely normative at the same time.

In what follows, we aim to restate the constitutive rules argument in such a way as to show that it is indeed especially pressing for inferentialists and that on their account of meaning it is extremely difficult to maintain the claim that meaning is genuinely prescriptive.

4. Ajdukiewicz's theory and constitutive normativity

In order to show that defining meaning in terms of inferential relations might quite easily lead to the conclusion that meaning is normative only in the constitutive sense, we will present the theory of meaning developed by Ajdukiewicz in the 1930s. Although this theory certainly differs in many respects from contemporary inferentialism, it shares many important affinities with the way Sellars conceptualizes meaning. These affinities are deep enough to make Ajdukiewicz theory a useful, albeit simplified model on how the idea of defining meaning in terms of inferential relations can lead to the conclusion that the only norms of meaning are constitutive ones. Ajdukiewicz theory differs from contemporary approaches because it deals with language understood as a strictly defined formal system, however, basic inferential ideas are already present in his system.

Ajdukiewicz developed his theory in two papers (Ajdukiewicz 1978a and 1978b). The crucial observation behind the theory is the question of how speakers of a given language settle semantic disputes. Ajdukiewicz pointed out that every now and then people start to suspect that their interlocutors do not use words the same way they do. What happens next is that the users retreat to a number of platitudes that every speaker of the language have to accept if they are to be counted as a speaker of this particular language. Prescriptions which point out sentences users have to accept in given circumstances are called "meaning directives".

In general, directives can be described as rules which instruct the user to accept a specific sentence in specific circumstances. Depending on the circumstances presented in a given directive, Ajdukiewicz differentiated between three types of directives: axiomatic, deductive, and empirical. To understand how they work within the theory, it is probably best to start with deductive directives. Consider a standard example of a Modus Ponens rule. A deductive directive associated with this rule is a prescription which states that whenever the users accepts a conjunction of an implication and its antecedent, they cannot refrain from accepting the consequent. If they fail to follow this rule, they will not be taken seriously by the community. They will either be seen as joking, provoking, or simply as someone who does not understand the meaning of the expressions they use. This example seems to be fairly intuitive because this is more or less how we normally learn logical connectives and test their understanding. Ajdukiewicz's ingenious idea was that similar rules enforce meanings of every non-compound expression in a language. In other words – if a language user wishes to be treated as a competent user of a given word, they have to act in accordance with the directives connected with this specific word and if they want to be treated as a competent language user they have to follow rules associated with a great deal (admittedly unspecified) number of words.

The other two types of directives are: axiomatic directives, which instruct the user to accept a sentence in every situation, and empirical directives, which instruct them to accept a sentence if they happen to have a certain sensory experience. A good example of an axiomatic directive is the rule which states that identity sentences such as $a=a$ are to be accepted in every circumstance. An illustration of an empirical directive proposed by Ajdukiewicz is a rather graphic example of a patient who should accept a sentence "It hurts!" when his tooth nerve is touched.

The theory does not tell us anything about whether the person "understands" the rule. Their task is only to act in accordance with it. It is also crucial to point out that the theory expects the users to react accordingly to directives whenever they are challenged by other community members. A user who accepts certain sentences does not have to follow every inferential pattern that exists in the language or inform the community about their every feeling or sensation. They only have to be disposed to do it whenever they are asked to. The way Ajdukiewicz's theory defines meaning of an expression is that it identifies it with the distribution or placement of this expression within the structure of all directives that contain it.

One very strong consequence of these definitions is that they connect meanings of expressions with the structure of the language they are part of. After all, the notion of a "distribution" or a "place" in the structure makes sense only if the structure in question is fixed. There is no sense in saying that two expressions have "the same place" if the structures in which they are embedded are different. The result of this is very counter-intuitive: because the meaning of every expression is tied strictly to the structure of the meaning directives, it changes whenever the structure changes. But the structure of language changes whenever a new term is introduced to the language. It is so because if the term is to have any meaning, it has to come bundled with some new meaning directives which fix this meaning. But once we add new meaning directives to a language, we inevitably change the structure of directives.

Ajdukiewicz himself was not concerned by this problem because he restricted his semantics to a very special type of languages which he called "connected and closed". The notion of "connectedness" of a language is rather easy to grasp. What it means is that the language does not contain any isolated parts, that is, every expression within it connects to every other expression via a chain of meaning directives.

The property of being "closed" is definitely much more contentious. In a nutshell, a closed language is a language which cannot be further semantically expanded – it is impossible to add new meanings to it. The reason why it is impossible is that all possible connections in the network of directives are already exhausted so the language achieves its full semantic potential. Because of this, every attempt to expand it with a new term ends up with the term either becoming synonymous with one of the existing expressions (as its meaning directives repeat some of the existing directives)

or the new language becoming disconnected (when the new term does not use any connections with older directives). A language which is not yet closed is called an "open language".

The most significant question, from our point of view, is whether meaning in Ajdukiewicz's theory can be seen as a normative notion. Superficially, it might seem so as there are semantic "rules", which might be violated. Upon deeper reflection, however, it turns out that the normativity in question is of a strictly constitutive kind.

The main question to be asked is that whether there is any sort of prescription involved in the notion of directive? From our perspective the answer is a flat no. This is because there is little room in Ajdukiewicz's framework for a notion of violating a meaning-directive. If we focus on the situation in which we are dealing with a closed language, there is little sense in which one can break a semantic rule. If one uses a certain expression in a way which cannot be accounted for in terms of the meaning directives, then the consequence for the speaker is just that they would be considered using the expression in question with a different meaning in the sense that they would use the expression with a different set of associated directives. This is a characteristic feature of constitutive rules: in a way it is impossible to violate them: when one signs the inappropriate form on the marriage ceremony, it is not that the marriage was started badly; there is no marriage at all.

Applying Ajdukiewicz's semi-formal apparatus to the situation, it might be said that a mere rejection of or a change in one directive from the set of directives associated with a given expression changes the structure of directives, and thus changes the meaning of the said expression. Moreover, as we are dealing with a closed language here, such a change results in a change of language.

Such a conclusion might seem very counter-intuitive, but it is worth bearing in mind that the concept of language Ajdukiewicz deals with is not folk but a highly technical one, which is chosen for specific theoretical purposes. If this notion is adopted, then it must be said, however strange it would sound, that there is nothing like a semantic mistake. When a speaker starts to violate the semantic directives, they simply start to use a different language (in Ajdukiewicz's sense of the term). Perhaps there are some situations in which a person does not speak any language – if their behaviour is impossible to be made consistent with any possible set of directives.

The principal problem here is whether users have any prescriptive reason to prefer one language to any other. Let us say that violating a certain directive would result in me ceasing to speak *Lx* and starting to speak *Ly*. Is there any prescription to follow *Lx*? It seems not – the change in meaning, in itself, has no normative import (some other users would for example stare at me, but this is not a semantic phenomenon according to Ajdukiewicz, but merely a pragmatic one).

The conclusion is that in Ajdukiewicz's theory directives play a distinctively constitutive role. A language is an abstract entity which is constituted by the totality of the meaning directives of all the expressions of this language. Semantic rules play the role, as it were, of classificatory devices, which allow us to say to which language a certain expression in a certain context belongs to (and, henceforth, which language the user is using). They are not semantic rules which are to be "followed" in a strong sense – they are not usually intentionally adopted and, more importantly, they do not create any genuine obligations for the users.

To sum up, Ajdukiewicz's theory gives an example of how one can treat meaning in inferential, anti-representational terms, and, at the same time, how one can treat semantic rules as purely constitutive ones, without claiming that there is any prescriptivity to it.

5. Contemporary inferentialism and constitutive rules

Contemporary inferentialism is obviously a very different theory than that of Ajdukiewicz. The main source of difference is the fact that, contrary to Ajdukiewicz, contemporary inferentialists aim at creating a theory which could be realistically applied to natural languages.

If one wants to create a feasible theory of meaning for natural language, then the concept of closed languages is of no use. This is because it is extremely implausible from the point of view of natural language analysis that a single change in one inferential rule, which co-defines meaning of one word, is enough to change the whole language. We naturally think of language as a dynamic system in which quite significant changes are possible. Thus, a concept of meaning that would have similar consequences to the Ajdukiewicz's theory would be blatantly inadequate.

However, the rejection of the idea of closed languages leads to a difficult question: how to combine, on the one hand, the idea that a meaning of an expression is somehow constituted by the assorted rules of material inferences and, on the other, the insistence that it's possible to change those rules while speaking the same language.

According to inferentialists, the central notion in this context is the notion of "similarity of meaning". This is especially important for Sellars (see e.g. 1973 and 1974). Sellars rejects the idea that we can talk about sameness of meaning in the strict sense, and, consequently, that the so-called conceptual change normally results in a complete change of meaning of terms involved in such a change (this is especially important in his discussion of theoretical terms). Williams summarizes his position:

> Since inferential engagements change over time and vary between persons, sameness of meaning is similarity of meaning, as Sellars is well aware. But similarity is always sufficient similarity for particular purposes. (Williams 2016, 250)

Since similarity of meaning is context-sensitive and not strictly defined, the inferentialist can allow for a slight change in meaning understood as set of inferential norms, without having to resort to the idea that each time such a change arises we deal with an entirely different concept. What we deal with is the same term with a slightly-different-yet-similar meaning.

The problem, however, remains, whether this change of focus – from sameness of meaning to similarity of meaning – weighs substantially on the relation between constitutive character of meaning and its alleged normativity. We believe it does not, and we are about to argue for this presently.

The main offshoot of our discussion of Ajdukiewicz's conception was that within its framework there is no such thing as genuine normativity of meaning. As rules of meaning are conceived in strongly constitutive sense, the result of "violating" meaning directives is that the speaker ceases to speak a given language *Lx* and starts to speak some other, albeit similar, language *Ly*. As languages are considered to be abstract systems of directives, there seems to be no prescriptive reason to prefer one language over another.

The question that arises now concerns the consequence of violating the inferential rule which is constitutive of meaning of a certain expression within the framework of contemporary inferentialism. Certainly, it is not the case that each time we violate an inferential rule we change the language – the framework of contemporary inferentialism allows us to violate the inferential rule and still speak the same language as before.

There are two possibilities of rule-violation: first is a simple, inadvertent mistake, like a slip of the tongue. The second is an intentional flout, when one deliberately violates the inferential rules of language.

Let us consider the second, theoretically more interesting option. Take a user of language who is fully aware that a certain inferential pattern is definitive of meaning of a certain expression and deliberately violates the norm (say, by stubbornly refusing to accept certain material inference). This, of course, is a thing that might actually happen, but, according to the strongly normativist view of semantic rules, we should be entitled to say that this person should not have done this; if the prescriptive account of semantic rules is on the right track, there is a sense in which this person should have used the word in accordance to the inferential rules that define the meaning of this word.

However, there is a strong worry that the inferentialist cannot really endorse such prescriptive claims. On the inferentialist account, what happens when a speaker uses an expression with a slightly different set of inferential patterns than we do, what we should really say is that this person uses this expression with a similar-yet-slightly-different meaning.

Is there anything "wrong" with using words with a similar-yet-slightly-different meaning? It does not seem that an inferentialist has any resources to make such a claim. Obviously, the behaviour of such a person could (and most likely would) be subject to some form of verbal correction from other members of the community, but the question is whether there is a prescriptive reason for the speaker to use the word according to communal standards. The mere fact that other people would have a tendency to correct the user provides in itself no prescriptive reason for the user in question to avoid behaviour leading to such a correction (this is stressed by Hattiangadi 2003 and Kaluzinski 2016). As Kaluzinski notes, grounding a notion of meaning in the idea of practice of making corrections might be easily taken to be a form of dispositionalism about meaning.

Such a dispositionalism – according to us – is by no means a normativist position. Rather, it is similar to what Kripke (1982) called "social dispositionalism" and rightly rejected as an inadequate solution to the problem of normativity of meaning.

This observation can be strengthened if we consider the problem in terms of practical reasoning. If we try to reconstruct the reasoning of a subject in such a social-dispositionalist framework, it would most probably look like this:

1. I want to avoid correction by the community;
2. In order to avoid correction by the community, I need to follow the socially accepted inferential rules associated with the expression I am using;
3. I should follow the socially accepted inferential rules associated with the expression I am using.

In such a reasoning, meaning does not play any prescriptive role – it only serves to delimit the options which are available to me, given the fact that I want to avoid correction. Should I, however, have no problem with being corrected, then – on the purely dispositionalist social account, I would have no genuine reason not to modify the existing inferential patterns and use the expressions with slightly-different-yet-similar meanings.

A similar diagnosis can be given in a situation where a subject makes an involuntary mistake, a semantic equivalent of a slip of the tongue. It might be truly said of such a person that they used the expression incorrectly, but it does not mean that there is any prescriptive semantic normativity involved. Again, if we tried to reconstruct the potential practical reasoning of the subject involved, it would have the following form:

1. I want to use the expressions the same way as the community;
2. In order to use the expressions the same way as the community, I need to follow the socially accepted inferential rules associated with the expression I am using;
3. I should follow the socially accepted inferential rules associated with the expression I am using.

Thus, when users who want to be faithful to socially accepted inferential norms find themselves making an unintentional mistake, there is a sense in which they should not have done so. However, this is contingent on their intention to follow socially accepted inferential patterns of use. Should they want to deviate from them, there would seem to be no reason for them to do so (apart from the previously discussed motivation to avoid correction).

To sum up, although there are important differences between the way contemporary inferentialists and Ajdukiewicz conceptualized meaning, these differences seem to have very little impact on the problem of normativity of meaning. There seems to be no way in which the change from the strict idea of closed languages to the liberal idea of similarity of meaning can contribute to the debate on prescriptivity of language. It seems that even within the liberal framework, the language rules serve simply as classificatory devices and not as prescriptive norms.

We think that what makes the inferential approach especially susceptible to the argument from constitutive rules is the fact that if one decides to define meaning in terms of a "correct" inferential relation between expressions, it might follow quite naturally that these inferences play meaning-constitutive role. Once one admits this, then it is quite hard to argue that there is a way in which prescriptivity can be read into meaning ascriptions.

It might be said that the line of reasoning presented in last two paragraphs is not convincing as it might be generalized too easily – is it not the case that any set of rules can be presented as an abstract set and thus it would seem as not normative?[2] We believe that this is not the case; Ajdukiewicz's framework indeed could be used to semi-formalize other systems of constitutive rules (although the usefulness of such a formalization is debatable). However, there is little reason to think that we could use Ajdukiewicz's model to show that systems of rules which is intuitively prescriptive would turn out to be constitutive. There seems to be little room to present e.g. rules of social etiquette or morality as rules of Ajdukiewicz-style system of directives.

To see this contrast, consider a following example. Picture a speaker at a funeral. In situation A she violates a semantic rule. There are different

[2] We are grateful to an anonymous referee for drawing our attention to this point.

ways in which her utterance could be explained (it may for example be seen as a result of her emotional state) but she will not be seen as "really saying" what she appears to be saying. We could say that her move in a language game will be cancelled because it will be ignored by the community. In situation B she violates an etiquette rule by using a curse word. Even though the psychological explanation used by the listeners to explain her violation may be the same, her act will not be cancelled or ignored. She will be accounted for making a bad (as opposed to impossible or wrong) move in a language game.

6. Possible reply

We believe there is a way in which an inferentialist can try to refute this argument. In the framework of contemporary inferentialism, meaning is treated as an essentially social phenomenon (see e.g. Brandom 1994, Peregrin 2012), and this social aspect of language is treated by far more seriously than in Ajdukiewicz's proto-inferentialism. For contemporary inferentialists, the fact that languages are social phenomena is crucial to the proper understanding of meaning.

For most contemporary inferentialists, the way of thinking about the social aspect of language, which we have presented above, namely the social dispositionalist, is thoroughly inadequate. The social dispositionalists see linguistic interactions from an impersonal, third-person, naturalistic perspective, in which the process of mutual corrections is described in purely non-normative terms. Such an outlook is obviously inadequate when it comes to explaining the normative aspect of language.

Instead of adopting a social-dispositionalist account, the inferentialists describe the social aspect of meaning in irreducibly normative terms. This allows them to look at the process of attributing correctness and incorrectness of linguistic utterances in normative terms from the very start.

How this connects with the problem of practical reasoning posed by Glüer and Pagin? In the social-normativist framework the proper account of the practical reasoning should look more or less like that:

1. I ought to use the expressions the same way as the community;

2. In order to use the expressions the same way as the community, I need to follow the socially accepted inferential rules associated with the expression I am using;
3. I should follow the socially accepted inferential rules associated with the expression I am using.

The first premise of such a reasoning has an explicitly prescriptive character, as the ought here is not taken to be an "ought" derivative of some practical interest, but rather as an expression of genuine social obligation. Thus, on such a construal, we might claim that meaning-ascriptions are "genuinely normative".

The question now arises, however, whether this normativity is a purely semantic one. We believe it is not. The first normative premise is not inherently tied with meaning taken as an abstract semantic notion, but with the social aspect of language. The first premise in this reasoning can be justified only if one accepts the premise stating that there is some prescription involved in the fact that the community uses a language in a certain way.

Such a thesis might be justified by resorting to some social norm – like the norm of solidarity or positive conformism – which dictates that a certain form of linguistic behaviour is to be normatively preferred to the other, namely, the behaviour which conforms to the inferential patterns accepted by the community should be preferred to the one which deviates from the socially accepted forms of use. Obviously, this is a *prima facie* defeasible norm – there are many reasons which could justify breaking actually existing linguistic rules (Whiting 2007 stresses the importance of the fact that semantic norms are *prima facie* in character). There might be moral, aesthetic, or pragmatic reasons for using a different set of inferential rules than the ones which are communally accepted. Nonetheless, the norm in question still holds, even if other norms can override it.

The point of contention between our account and the one which seems to be endorsed by inferentialists is that for us it is incorrect to say that the normativity we are dealing with here is a distinctively and exclusively semantic one. There is a sense in which prescriptive normativity enters semantic discourse, as envisaged by inferentialists, but this is a social normativity and not a distinctively semantic one.

The most important lesson from the discussion of Ajdukiewicz's theory is that it is possible to adopt a sterilized version of inferentialist theory of meaning, one that abstracts from the socio-normative aspect of language. In such a framework, meaning is "normative" in a purely constitutive sense. Only when we adopt a social-normative outlook do the ascriptions of meaning become saturated with prescriptive normativity.

This diagnosis explains, in our opinion, two things. First, it shows why proponents of inferentialism treat the thesis of semantic normativity as something which is uncontroversial within their theory: this is because for them the socio-normativist account of linguistic practice is something that goes without saying. It also explains why the idea that inferentialism leads to normativism might be easily challenged: what makes meaning normative is not the fact that it should be defined in terms of material inferences but the fact that language is a social phenomenon and this social aspect of language should be accounted for in normative terms.

Such a theory of sources of normativity of meaning also provides a convincing reply to an old challenge to the idea of semantic normativity posed by Davidson (2005), who complained that the proponents of normativity of meaning make an absurd claim that people might be "obligated to a language" (Davidson 2005, 118). In the socio-normative model, there are no obligations to a language understood as a system of abstractly conceived rules but there are indeed obligations towards a community which uses words according to certain inferential patterns and these obligations do not boil down to mere pragmatic interests (like the need for a smooth communication).

To sum up: in the framework of contemporary inferentialism, meaning is indeed a normative notion but only when we look at language from a socio-normative perspective. If we take meaning to be determined solely by abstractly understood inferential norms, then the normativity of meaning is of a purely constitutive kind.

This conclusion might seem slightly catholic, but it has potentially important consequences. On our take, semantic normativity is derivative of social norms. If this is actually the case, then one cannot hope to ground social normativity in the semantic one – and such hope seems to be implicit in some inferentialist writings. However, if our reasoning is correct this cannot be achieved.

7. Conclusion

Ajdukiewicz's theory provides a useful model of how one could build a theory of meaning that would define the notion in terms of inferential relations and abstract from prescriptive social normativity. Although language might be described in normative terms the norms in question are constitutive rules. In this respect abstractly conceived semantics resembles chess more than ethics. This shows that the very idea of defining meaning in terms of inferential relations does not lead in itself to any form of strong normativist approach to meaning. If one accepts Peregrin's idea that normativity of meaning is definitional of inferentalism, such a conclusion might look like a reductio ad absurdum. However, this is not the case – the right conclusion is that the normativity of meaning which is in play in inferentalism need not be of the strong, "objectively prescriptive" variety, even though most of inferentialists have seemingly assumed it to be such.

Still, the strongly normativist approach to meaning might be justified within the inferentialist framework, but only when the social aspect of language is taken into account, and this social aspect of language is accounted for in normative terms from the very start. This shows that within the framework of contemporary inferentialism, prescriptive normativity of meaning should be treated as derivative of prescriptive social normativity, and thus semantic normativity cannot be treated as basic and grounding other forms of normativity.

Acknowledgements

The first version of this paper has been presented at the Why rules matter? Workshop in Prague, November 2 – 4 2016. We would like to thank all the participants for the stimulating questions, and, especially, Jaroslav Peregrin and Hans-Johann Glock for their critical encouragement. Moreover, we would like to thank Tadeusz Ciecierski, Michał Zawidzki, Iza Skoczeń, Zuzanna Krzykalska, Paweł Banaś, and Bartosz Janik, as well as an anonymous referee for their helpful remarks. The work on this paper was funded by National Science Center, Poland, grant under award number UMO-2014/15/B/HS1/01928.

References

AJDUKIEWICZ, K. (1978a): On the Meaning of Expressions. In: Giedymin, J. (ed.): *Kazimierz Ajdukiewicz, The Scientific World-Perspective and Other Essays 1931-1963*. Synthese Library vol. 108. Dordrecht: D. Reidel Publishing, 1-34.

AJDUKIEWICZ, K. (1978b): Language and Meaning. In: In: Giedymin, J. (ed.): *Kazimierz Ajdukiewicz, The Scientific World-Perspective and Other Essays 1931-1963*. Synthese Library vol. 108. Dordrecht: D. Reidel Publishing, 35-67

BOGHOSSIAN P. (1989): The Rule-Following Considerations. *Mind* 98, 507-549.

BOGHOSSIAN, P. (1993): Does an Inferential Role Semantics Rests upon a Mistake? In: Villanueva, E. (ed.): *Philosophical Issues* 3, 73-88.

BOGHOSSIAN, P. (2005): Is Meaning Normative? In: Beckermann, A. & Nimtz, C. (eds.): *Philosophy—Science—Scientific Philosophy*. Paderborn: Mentis.

BRANDOM, R. (1994): *Making it Explicit*. Cambridge, Massachusetts: Harvard University Press.

BRANDOM, R. (2000): *Articulating Reasons*. Cambridge, Massachusetts: Harvard University Press.

BURGESS, A. (2015): An Inferential Account of Referential Success. In: Gross, S., Tebben, N. & Williams, M. (eds.): *Meaning Without Representation*. Oxford University Press.

CHRISMAN, M. (2010): Expressivism, Inferentialism and the Theory of Meaning. In: Brady, M. (ed.): *New Waves in Metaethics*. Palgrave Macmillan.

DAVIDSON, D. (2005): The Social Aspect of Language. In: *Truth, Language and History*. Oxford: Clarendon Press.

GLOCK, H.-J. (2005): The Normativity of Meaning Made Simple. In: Beckermann A. & Nimtz, C. (eds.): *Philosophy—Science—Scientific Philosophy*. Paderborn: Mentis.

GLÜER, K. & PAGIN, P. (1998): Rules of Meaning and Practical Reasoning. *Synthese* 117(2), 207-227.

GLÜER, K. & WIKFORSS, Å. (2016): The Normativity of Meaning and Content. In: Zalta, E. (ed.): *The Stanford Encyclopedia of Philosophy (Spring 2016 Edition)*, URL = https://plato.stanford.edu/archives/spr2016/entries/meaning-normativity/.

HATTIANGADI, A. (2007): *Oughts and Thoughts*. Oxford: Clarendon Press.

HATTIANGADI, A. (2003): Making it Implicit: Brandom on Rule-Following. *Philosophy and Phenomenological Research* 66(2), 419-431.

KALUZIŃSKI, B. (2016): Assessment, Scorekeeping and the Normativity of Meaning: a Reply to Kiesselbach. *Acta Analytica* 31(1), 107-115.

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

MCDOWELL, J. (1984): Wittgenstein on Following a Rule. *Synthese* 58, 326-363.
MILLER, A. (2010): The Argument From Queerness and the Normativity of Meaning. In: Grajner, M. & Rami, A. (eds.): *Truth, Existence and Realism*. Fankfurt: Ontos.
MILLER, A. (2012): Semantic Realism and the Argument from Motivational Internalism. In: Schantz, R. (ed.): *Prospects for Meaning?* Berlin & Boston: De Gruyter.
PEREGRIN, J. (2012): Inferentialism and the Normativity of Meaning. *Philosophia* 40(1), 75-97.
PEREGRIN, J. (2014): *Inferentialism: Why Rules Matter*. Palgrave Macmillan.
SEARLE, J. (1969): *Speech Acts: An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press.
SELLARS, W. (1954): Some Reflections on Language Games. *Philosophy of Science* 21(3), 204-228.
SELLARS, W. (1973): Conceptual Change. In: Pearce, G. & Maynard, P. (eds.): *Conceptual Change*. Boston: D. Reidel, 77-93.
SELLARS, W. (1974): Meaning as Functional Classification. *Synthese* 27 (3-4), 417-437.
SHAPIRO, L. (2004): Brandom on the Normativity of Meaning. *Philosophy and Phenomenological Research* 68(1), 141-160.
SMITH M. (1994): *The Moral Problem*. Oxford: Basil Blackwell.
WHITING, D. (2007): The Normativity of Meaning Defended. *Analysis* 67(294), 133-140.
WHITING, D. (2015): Conceptual Role Semantics. In: Fieser, J. & Dowden, B. (eds.): *The Internet Encyclopedia of Philosophy*.
WIKFORSS, Å. (2001): Semantic Normativity. *Philosophical Studies* 102(2), 203-226.
WILLIAMS M. (2013): How Pragmatists Can Be Local Expressivists. In: Price, H. (ed.): *Expressivism, Pragmatism and Representationalism*. Cambridge: Cambridge University Press.
WILLIAMS M. (2016): Pragmatism, Sellars, and Truth In: O'Shea, J. (ed.): *Sellars and His Legacy*. Oxford: Oxford University Press.

Organon F 25 (2) 2018: 196-214

Reflected View on the Personal Afterlife

DANIEL KRCHŇÁK[1]

ABSTRACT: In this paper, I try to argue that, from the methodological position of reflected equilibrium, it seems to be reasonable to build a theory of personal identity that enables a person to continue her existence after the biological death of her body. This conclusion is supported by the argument that our practice reflects that our identity-presupposing concerns reach beyond biological continuity. We have also good reasons to maintain such concerns and practices. As the best candidate to implement such concerns in a theoretical account of practical identity, I will identify the person-life view, where personal identity depends to a great extent on social conditions. I also show how this theory can implement the classical belief in the afterlife, and how it could conceptualize the difference of the afterlife from a physicalistic and a theistic point of view.

KEYWORDS: afterlife – Marya Schechtman – reflected equilibrium – Pascal's wager – personal identity – person-life – Radim Bělohrad – Samuel Scheffler.

1. Introduction

"*There are no words to describe the bravery required to take such an action. ISIS were robbed of a predictable macabre propaganda opportunity by Ryan's action. I personally believe he deserves the*

[1] Received: 30 January 2018 / Accepted: 24 April 2018

✉ Daniel Krchňák

Department of Philosophy
Faculty of Arts, Masaryk University, Arna Nováka 1
602 00 Brno, Czech Republic

e-mail: 362596@mail.muni.cz

> *very highest of military honors for such outstanding bravery in the face of such a barbaric enemy*." Mark Campbell[2]

This quote doesn't seem too interesting, philosophically, at first glance. The point is that the celebrated act was an act of suicide undergone to prevent being taken as a prisoner by ISIS. That means that Mark Campbell (Kurdish rights activist) claims, that Ryan Lock deserves some honors, though he knows that he is dead. This is an example of ascribing personally relevant concerns to a person, who is biologically dead. In this paper I will try to think through possible reasons and theoretical consequences of such a common practice (I will try to show, that it is common practice indeed as well). Many traditional or classical theories allow that we continue to exist after our biological death. However, the continuity of personal existence rests here on the presupposition of conscious experience after death. Since the question of whether there is such an afterlife is highly controversial, the argument will be made without that assumption, and the idea of a traditional afterlife will be revisited after the argument is made. I will try to argue that when we apply the methodology of reflective equilibrium, it seems to be reasonable to strive to build a theory of a personal identity which allows for the person to continue after the biological organism ceases to exist, even if we don't accept the continuity of experience after death.

2. Methodology

At first it is nevertheless crucial to make clear on which methodological steps the conclusion essentially rests. There is a strong tradition in the area of philosophy of personal identity which builds theories of person in purely metaphysical terms. Both Derek Parfit and Eric Olson as the main figures of the most influential – psychological and animalistic – strands deliberately try to make such an account, which does not consider our everyday practice in the first place. Mark Johnston (1997) goes still further and claims that there is no relation between our practical concerns such as moral responsibility, compensation, survival or self-concern,

[2] Quoted from Robson & Wheatestone (2017).

which are traditionally held to presuppose personal identity (I will call such a concerns in accordance with Bělohrad (2016, 8) "i-concerns"), on one hand and the concept of person on the other hand. A crucial disadvantage of such an approach is that this prevents us from the possibility of reforming our practice. Such theories don't take in account how (psychologically) deep some practices and i-concerns lie. This easily results in a theory implying such grave practical changes (when it has some practical aspiration at all) that it doesn't have great chances to be effectively adopted (Bělohrad 2016, 51).

On the other hand, there is also a host of authors who begin their research in personal identity with the i-concerns (e.g. Schechtman 1997; Korsgaard 1989; Mackenzie & Atkins, 2008). Nevertheless, it isn't clear in which way this approach has a better position to bring a practical impact. Must this not remain a purely descriptive project? Bělohrad (2016) in this context suggests to apply the method of reflective equilibrium, which I will embrace here. I believe that the core of this method, which is widely spread in other areas of philosophical research (e.g. ethics, logic), is succinctly expressed in the insight of David Lewis:

> One comes to philosophy already endowed with a stock of opinions. It is not the business of philosophy either to undermine or to justify these preexisting opinions, to any great extent, but only to try to discover ways of expanding them into an orderly system. (Lewis 1998, 99)

This method gives us the possibility to reject belief that is a) not in accordance with our other beliefs and b) for us not more important than the sum of our beliefs which it contradicts.

The second important methodological step is bound to the acknowledgment of the plurality of i-concerns. For some i-concerns, we need a more narrowly defined entity to be able to apply these specific i-concerns. For example regarding anticipation, the future entity needs to be conscious, whereas for moral responsibility arguably it does not (Shoemaker 2007; see also Bělohrad 2016, 225). Now comes the question: "What happens when not all i-concerns are applicable?" There seems to be an agreement that the person exists when there is an entity which is subject to at least some of the i-concerns. In that manner, Schechtman (2016) defines *person* very broadly as an entity with person-life (the condition is so loose that,

e.g., people in pervasive vegetative states fulfill that condition). She is aware that some objects which fulfill her definition of a person are not able to engage in the full range of i-concerns and related practice, but that does not mean that this entity isn't a person at all. Similarly, Bělohrad (2016, 225) considers the human organism as the main entity relevant to i-concerns, though he is aware that for some kinds of i-concerns it isn't a sufficient condition. The reason for this step seems obvious. When we take i-concerns seriously as concerns that presuppose identity, we can conclude that where there is some i-concern, there must be a person (leaving aside the possibility that other i-concerns may not hold).

With this methodological background (which is admittedly rather unconventional, but also not completely novel), I will try to show that there is a possibility to argue for the continuing existence of a person after death even without any supernatural intervention or non-naturalistic occurrences. To my best knowledge, for the existence of personal afterlife was argued so far only from the theistic perspective. Though I am not sure that other authors embracing the reflective equilibrium method aren't already on the point of accepting my conclusions, none of them has directly addressed this topic, so even if it is a relatively evident implication of this method, I believe it is meaningful to explicate it.

3. I-concerns beyond the biological continuity

The first necessary step towards the desired conclusion is to show that we hold at least some i-concerns that go beyond the point of death (I will call these concerns *afterlife i-concerns*). I will focus on two of them that belong to the most important and most discussed – egoistical concern and compensation.

Self-concern (sometimes also called egoistical concern) is a special kind of practical concern which I feel exclusively toward my own person. I can be deeply concerned for my close ones and the concern for others could be even in some respects stronger as for myself, but egoistical concern is qualitatively different from the concerns we feel toward others. As the pain of others is phenomenologically different from the pain I personally experience, also the expected pain of others is different from expected pain that I personally will have to undergo.

I identify three institutionally supported ways in which egoistical concern goes beyond our biological death. Firstly, we have an afterlife self-concern for our bodies. There is a difference between our concern for corpses of others and concern for our own corpse. When we imagine that our corpse will be treated in some disrespectful way, we feel that it would be personally offending. On the other hand, when we imagine that it will be treated reverentially we feel honored. And also when we treat some corpse in some reverential way we are convinced that in that act we honor the person of the dead body. Afterlife self-concern is probably manifested most strongly in our conviction that we have a right to decide what should be done with our body after our death. That is not merely an airy intuition of a few people; this judgment is also reflected legally. At least in many countries, everybody has a right to decide whether their organs could be taken for transplantation or not.

Something similar applies to our material property. We are personally concerned about the question what will happen with our property when we die in virtue of being our property. We feel that our personal right is violated when we cannot control what will be done with our property after our death, and we have indeed a legal right to determine it by writing a will.

Thirdly, we have an afterlife self-concern in respect to our reputation. We feel the same kind of outrage when we imagine that someone will spread lies about us after our death, as if he were to do so while we were alive. There is again also a legal right to defend one's post-mortem reputation (through the relatives).

The second i-concern that I argue for, which goes beyond biological death, is compensation (in a broad sense that involves not only material compensation but also praise, for example). We tend to say that someone deserves compensation for what he has done, even though he is dead. The perfect example for that is the example mentioned at the beginning of my paper. Mark Campbell obviously doesn't see any problem in saying that Ryan Lock deserves honors, though he is dead. Again, there are also legal cases that are underlined by afterlife compensation judgments. Copyrights are a form of compensation for the effort of creating a certain product of which others can take advantage. Inheritance of copyrights (which is legally guaranteed) could then be seen as a post-mortem compensation for that effort. But there is also a legal right to be compensated for events that

happen after the biological death. The action for the protection of personality guarantees that slander and other kinds of reputational harm will be compensated for even when they happen after the biological death of the person. That is the reason why the deceased journalist Ferdinand Peroutka has the right to be compensated (in the case that the article that Zeman claims Peroutka has written was not written) for the words of the Czech president Miloš Zeman.[3]

4. The importance of afterlife i-concerns

I believe I made a point that we do have some afterlife i-concerns. I can now apply the second methodological step and say that, as we have beliefs in afterlife i-concerns, we are bound to belief in some form of existence of person after death. It seems to be very strange to believe that someone has a right to be compensated and at the same time to believe that he doesn't exist anymore. But there is still the possibility, that we should sacrifice our belief in the appropriateness of afterlife i-concerns in order to keep some other potentially more salient beliefs. To evaluate this possibility we first have to consider the psychological value of our afterlife i-concerns.

4.1. The value of collective afterlife

The first source, which gives us the possibility to appraise the importance of this belief, draws on the thoughts of Samuel Scheffler, summed up in the book *Death and the Afterlife* (Scheffler 2013). Here, Scheffler presents thought experiments that are intended to show that our values crucially depend on our beliefs concerning the fate of mankind after our (biological) death. In the doomsday scenario (Scheffler 2013, 18-19) we are invited to imagine how we would react emotionally, if we found out that 30 days after our own death, the Earth would be destroyed in a collision with a giant asteroid. Scheffler supposes that most of us would react with "profound dismay" (Scheffler 2013, 21), and that lot of things that we value would lose their value for us. There is a type of

[3] See the details of this famous affair on *Kauza Hitler je gentleman* [*The Affair: Hitler is Gentleman*] in Wikipedia: The Free Encyclopedia, retrieved online [05.05.2017]: https://cs.wikipedia.org/wiki/Kauza_Hitler_je_gentleman.

activity in which this is quite clear. The value of all projects, where a) the ultimate success is perceived as laying in the distant future or b) the value of the project derives from the benefits for a large numbers of people over a long period of time, is obviously threatened (Scheffler 2013, 24). A paradigmatic example is cancer research or improving the social institutions. But the novel by P. D. James *The Children of Men* suggests that also far more routine aspects of our lives would be threatened (James 1992, 38). One way how Scheffler explains this supposed reaction is by noticing "something approaching a conceptual connection" (Scheffler 2013, 60) between valuing something and wanting something to be preserved. "To value X is normally to see reasons for trying to preserve or extend X over time" (Scheffler 2013, 60). When we know that the Earth would not be preserved, we would know as well that the things that we value would not be preserved. So as long as we are valuing anything (except for quite few exceptions), it is important for us to know that when we die everything else stays quite the same.

As one of the most valuable things for us are our personal relationships, it is both very important and desirable for us that there remains a network of valuable social relationships after our death, out of which we are wrenched. In this respect, it is more important for us that our close ones survive than that we personally survive. Through the survival of other persons, we can still retain a "social identity". According to Scheffler, many people seem to feel that "not being remembered is what being 'gone' really consists in" (Scheffler 2013, 29-30). When you know that some people who value their relationship with you stay after your death it makes you feel that you have a place in the social world of the future. On the other hand, when this is missing you are faced with the frightening prospect of a blank eternity of nonexistence. Scheffler identifies this as a powerful imperative for those who are bereaved to not forget.[4]

One can expand or specify this imperative not to forget to the larger scale of practices that help to keep the social identity. These practices include, I believe, the range of afterlife i-concerns that I discussed earlier. Scheffler unfortunately doesn't specify what he means by the term "social identity", but he describes this kind of concern also with the term "personalized relationship to the future" (Scheffler 2013, 31). Here it is quite clear

[4] The whole paragraph paraphrases Scheffler (2013, 29-30).

that when I do have a right to write my will, I feel personally more involved in the world that comes after my death. The same applies when I know that there will be some legislative power to protect my reputation, or that my dead body will be treated with some respect.

In the previous paragraphs, I showed that there is arguably an important psychological link between valuing things and belief in the collective afterlife and that the value of collective afterlife rests to a great extent on its ability to create a personalized relationship to the future or the social identity after our death, which in turn seems to rest to a great extent on practices bound to our afterlife i-concerns. This means that the rejection of these practices could induce not only a bigger fear of one's own biological death, but it could affect our valuing of things in general.[5]

4.2. À la Pascal's Wager

Another line of argument about reasonableness of afterlife i-concerns doesn't support the thesis about the importance of our post-mortem i-concerns for us, but presents a reason for keeping such i-concerns besides the importance for us as living beings on this Earth. The argument takes a form similar to the famous Pascal's Wager. The first premise says that we are not sure, whether there is some life after death. Though some think that it is completely impossible that we could survive our death (e.g. Johnston 2010), there are lot of models that defend the position that it is at least logically possible to survive the biological death.[6] And there is arguably also *some* empirical evidence for the existence of the afterlife – for example in the area of near-to-death-experience or parapsychology (Hasker & Taliaferro 2014). So I suppose it is quite safe to present our situation as an agnostic one. In this position of uncertainty we have the possibility to act as if the biologically dead persons would continue to exist, or as if they would cease to exist.

---

[5] Scheffler in his text rejects personal survival after biological death, but his distinction between collective afterlife and personal afterlife rests mainly on the personal survival (see Scheffler 2013, 65), which my conception doesn't necessarily entail. My disagreement with Scheffler is insofar just terminological.

[6] For example simulacrum model, falling elevator model, constitution account etc. See Green (undated), Hasker & Taliaferro (2014).

Let us first take a look on the possibility that there is no afterlife in the ordinary sense. In the previous arguments I argued that there is a great advantage (for the living persons) to act as if the biologically dead persons would continue to exist. But for the sake of this argument I could even admit that there are possible costs for such a behavior. When we act in this way, we arguably lose the opportunity to transplant organs from bodies of those who reject it and the possibility to ignore the testaments of rich people who want their rich relatives to inherit their belongings and possibly distribute the heritage among the more needy (though there is a worry how this could work in practice). So I can admit that when there is no afterlife in the classical sense there are some costs of acting as if the dead one would continue to exist, but these don't prevail over the benefits.

The situation changes dramatically when the dead persons continue to exist (in the ordinary sense). In such a case, when we act as if they don't exist anymore there is at least some probability that they are harmed through our behavior, that they could feel offended or hurt by our behavior. They would probably feel in a similar way as when a friend doesn't want to be one's friend any more without any appropriate reason. The described situation is schematically presented in the following table.

	Continuing to exist (and care about our world)	Ceasing to exist/not care
Acting as continuing	C&B + Benefits for survivors	C&B
Acting as disappeared	C&B + Costs for survivors	C&B

Table 1: Afterlife wager
(C&B stands for costs and benefits)

4.3. Advantages of believing in the classical notion of afterlife

So far, I have only addressed the possibility of afterlife shared by those who don't believe in any biological or soul-like continuation of the person after death or are at least agnostic about it. But it seems to be relevant to also highlight the special psychological advantages of believing in the

afterlife in the classical theistic sense. Scheffler identifies four main features of the importance of the traditional afterlife. Firstly, it simply allows personal survival and reduces the fear of death. Secondly, it offers the possibility of reuniting or at least communicating with loved ones. Thirdly, it allows to believe in some kind of cosmic justice. It offers the possibility of appropriate afterlife compensation for all the terrible suffering, or afterlife punishments for the most grievous wrongdoings.[7] Fourthly, it gives life its cosmic meaning. It seems at least possible to argue that if there is no afterlife then nothing ultimately matters. By this point, however, Scheffler argues that life without classical afterlife apparently doesn't lead to life without meaning in reality. Many people live life without this belief and it seems that it doesn't diminish the extent to which things matter to them and they are engaged in a "full-array of valued activities and interactions with others" (Scheffler 2013, 71).[8]

Nevertheless, this seems to be a problematic statement. It is quite clear that such people can't see the same meaning in, e.g., prayer for the dead or in attempts to communicate with the loved ones. But given that there isn't any cosmic justice, it seems to be clear that it changes the value of moral behavior as well. It is relevant in this context to mention the story of the philosopher Holm Tetens, who after many years of being atheist/agnostic about the classical afterlife converted to theism/belief in the classical afterlife (Tetens & Scholl 2016). In his book in which he tries to rationally defend the theistic belief, he argues that given (at least) the uncertainty about the truthfulness of naturalistic explanations of the world,[9] it seems to be reasonable to choose such a metaphysics, which allows us to avoid prob-

---

[7] The possibility of punishment is not explicitly mentioned by Scheffler but, for example, Scholl admits that, for him, this is the most attractive aspect of afterlife (Tetens & Scholl 2016; 57:25-57:35).

[8] Nevertheless, this stance seems to be vulnerable to the following objection. One can imagine that in the same manner as one got used to the idea of the non-existence of the classical afterlife, one could become accustomed to the belief in the non-existence of collective afterlife. Though there may be satisfying answers to this objection, my claim seems to be less vulnerable, if I claim that those who don't believe in classical afterlife miss some motivation to engage in some projects.

[9] Tetens stresses in the first place the inability of explanation of the mind/body problem in naturalistic terms (Tetens 2015).

lematic attitudes in the fight against evil and suffering. Because the naturalistic view that excludes a classical afterlife presses us to adopt such a problematical stance as a "resignation, tragic opposition, cynical egoistical hedonism or the self-destroying delusion of self-redemption and in every case a moral awkwardness, giving a meaning of great amounts of evil and suffering in the best case as a mean to a human progress" (Tetens 2015, 78) it is more reasonable to adopt the theistic-redemptionistic metaphysics, which allows to avoid such an attitude. So, for Tetens, the promise of redemption, vindication and justice in the coming world presents a deciding reason to adopt a new whole ontological frame. That seems to be an evidence that, for some of us, the perspective of classical afterlife offers us still lot more than the perspective of the collective afterlife.[10]

To sum up my argument so far: I have argued that some of our i-concerns reach beyond biological death. I also showed that we have some quite important reasons to preserve these i-concerns. In order to be able to say whether it is really reasonable to embrace a belief in afterlife, now we have to look for a theory of personal identity that could implement naturalistic afterlife and consider which other beliefs such a theory forces us to sacrifice.

5. Person with an afterlife

A model that is the most frequently associated with the possibility of afterlife is the dualistic model of body and soul. The presented argument maybe gives some more attractiveness to this model, but I do not believe that the importance of afterlife i-concerns is powerful enough to overcome

[10] It might be nevertheless objected that, though belief in classical afterlife could bring a personal gain, it could be unfavorable for the society. A believer doesn't have such a big motivation to restore the righteousness on the earth as an unbeliever and so he could be more comfortable with, e.g., oppressive political conditions. From the position of Tetens one could reply that our capabilities to establish a righteous society are so negligible (in the face of the overwhelming power of unjustice) that there is not much sense in even trying to make a change. When we on the other hand believe that there is a real possibility of the victory of justice, we could have a lot more motivation to commit ourselves to some specific policies.

the deep ontological disagreement of materialists about dualistic metaphysics (though we saw in Tetens that one could be willing to radically change one's metaphysical framework in order to have the possibility of a classical afterlife). There are other accounts of surviving biological death that are materialistic (one could for example make use of psychological theory of personal identity; see Zimmerman 2013), but nevertheless presuppose the existence of God, which is again a problem of deep metaphysical disagreement. There is yet another account of survival of the biological death presented in Mark Johnston's *Surviving Death* (2010) which is built on naturalistic assumptions. This account is nevertheless not fitting in my argumentation insofar as it claims that the self is only an illusion. Johnston claims that the possibility of continuity after one's death lays in the redirection of our concerns to take a form of radical altruism (or *agape).* Through overcoming one's egoistical concerns and identification with humanity "one quite literally lives on in the onward rush of humankind" (Johnston 2010, 49). So it seems that this account, though it might have much in common with my arguments, isn't the best option to promote one's personal afterlife i-concerns.

5.1. Person-life view

The account that I see as the most promising for a theoretical anchoring of afterlife i-concerns comes from the book *Staying Alive* written by Marya Schechtman (2014). The theory described in this book, called person-life view (PLV), claims that a person is defined by living a person-life. According to Schechtman, it isn't accurate to think about the person as exclusively a forensic object. The forensic capacities (such as moral responsibility) are in our lives inseparably intertwined with all other activities (such as eating, sleeping, reproducing etc.).[11] There are – according to Schechtman – three different interrelated layers of a typical person-life (as a whole): a) individual attributes (biological and psychological), b) social interaction and c) social and cultural infrastructure (institutional

[11] As an example, Schechtman presents a situation of a wedding celebration, where eating and mating and traditions and rituals are all mixed together. It is not that we eat and mate and aside of it we also have traditions and forensic interactions (Schechtman 2014, 119).

framework of person-practices) (Schechtman 2014, 112-113). But according to PLV, all the features of the typical person-life needn't be present for a person to exist. In this sense, the concept of person is protean. The person-life is a cluster concept in a similar way as Chiong's (2005, 25) concept of biological life. Schechtman shows that even when there are attenuated individual attributes as in the case of a baby, a mentally handicapped person or a person in a pervasive vegetative state, there is still a whole range of person related actions of the people from their surroundings as well as legislative and cultural norms to treat such people in the person specific way, that enables the person to continue to exist (Schechtman 2014, 120).

On the other hand, Schechtman argues that all kinds of oppression and mistreatment (such as slavery) don't express that the oppressor treats the oppressed in a non-person-related way. It differs qualitatively from the way one would treat animals. For example, Slave Codes, which prevent slaves from testifying in courts, making contracts, buying or selling goods, etc., show acknowledgment of the slave's ability to do such things, which differentiates them from animals and other non-persons. Person specific treatment doesn't mean good treatment. In that sense even the oppressive institutional framework gives a human being a person-space (Schechtman 2014, 127).

But someone could object that we treat also other objects than people – typically pets – in a person-specific way. "There are pampered poodles, for instance, who wear sweaters and jewels, sleep in beds, have their births registered, go to doggie daycare and on playdates, are given therapy if they demonstrate anxiety, and eat 'people food' off of plates" (Schechtman 2014, 121). Schechtman points out here again that the attitudes that we hold toward pets qualitatively differ from those we hold toward a mentally handicapped child for example (though their mental forensic abilities could be at the same level). That seems to be apparent in the difference of the reactions of the parents of a cognitively disabled child, when they are confronted with the realization that their child won't be able to talk, to dress or feed herself on the one hand, and the owner of a pet, who gets the same information about her beloved poodle. In the first case we expect big emotional reaction, while in the second case we would expect puzzlement about expressing such a trivial statement. We are aware that children are not able to be included in all i-concern related practices, but that doesn't matter

because we expect them to be able of that in the future or in the past. But even if in some concrete case these expectations were irrational that wouldn't change anything. Probably no one thinks that the status of a child (in the sense that it would be no longer a person) should suddenly differ from other children just because its expected development is different. These children are the right *kind* of entity (because they are humans) and that is enough. This step should be easily seen as an expression of speciesism, but Schechtman tries to show that this step is not as arbitrary as it can seem at first glance. We have a deep natural tendency to treat other humans as persons, because they have through their biological outfit the best conditions to live with us in one community of persons – they are

> born from us (and later can reproduce with us) […], require the kind of nourishment and temperature regulation that are optimally provided by a human mother. They have the same sleep cycles we do, are nourished by the same foods, rely on the same senses, are subject to the same illnesses, can move at roughly the same speed, and so on. These are all facts about our biology, but they are also facts with immediate and wide-reaching implications for how we can and do live together. (Schechtman 2014, 124)

I didn't present entire PLV theory (with all its metaphysical consequences), but it seems clear to me that the main difference between PLV and other theories of personal identity lays in the crucial importance of the social aspect. Though other practical accounts such as the narrative theory put some weight on it,[12] no other theory I am aware of states social feature as a constitutive feature of a person. This seems to be very favorable for my purposes – my main concern is after all the (social) practice of backing up the personal identity in the first place.

Though Schechtman didn't comment in her book on the possibility of afterlife, it was objected that her theory doesn't, at least in principle, exclude this possibility (see Bělohrad 2014, 576). If we take features of person-life at face value, it seems clear that there are social interactions and

[12] For example, in her earlier narrative self-constitution view, Schechtman claimed that identity constituting narrative has to cohere with the beliefs about the most basic features of reality of other persons (Schechtman 2014, 119).

cultural infrastructure that are identity-relevant. There seems to be no difference in kind between the relationships or the legal rights and other social institution we have towards people in a pervasive vegetative state (which Schechtman claims to be identity-constituting) and towards people who are biologically dead. It was seen as a problem of the theory. Bělohrad (2014, 577) writes that "no one would accept that persons are entities […] that can survive their death, burial or cremation and that stop existing gradually as their position in person-space slowly disappears as their close friends and family forget them". I don't claim that people, who don't believe in the classical afterlife believe in surviving death (and insofar I agree with him). However, as I have already mentioned, Scheffler pointed out that being forgotten is (at least for some people) what being gone (or being no more) really consists of. Insofar, I believe it is not that unreasonable to believe that one stops existing gradually, as one is being forgotten by close friends and family.

6. Problem of "social afterlife"

The apparent problem for including the afterlife existence in the theory is that it radically amplifies the conventional nature of personhood. Even when we accept that it is not arbitrary – that human beings are treated by other human beings as persons – the existence of a person is (at least in some cases) determined by the contingent fact of the strength of the dying person's social network or by the number of people who will remember him. On this account, "immortality" really is gained through some history-changing deeds. The glory would really purchase long life for oneself, in a more literal sense than we are ready to agree on.[13]

One possible answer to such a question is that the existence of a person is still not completely conventional, because it is still parasitic on the

[13] It would also mean that even morally bad deeds can guarantee you longer life (through a herostratic fame). That seems to be morally problematical but only insofar as one presupposes that every form of existence is better that non-existence. I believe that it's arguable that such a kind of existence, which consists only of perpetual blaming, isn't worth striving for and therefore it is more like a kind of punishment to exist in such a form.

biological and mental outfit (if there were no forensic capacities by humans as a kind, there would be no human persons). Another possible answer seems to be at hand when we come back to the list of main i-concerns. As already mentioned, I don't claim that those not believing in a classical afterlife judge death as enabling personal survival. It could seem strange to claim that a person continues to exist but doesn't survive, but that is what we can claim given the pluralism of the i-concerns. In this line of argument we can claim that social interactions and social framework enables only personal existence without survival, and that means a very limited form of personal existence, which doesn't seem so counter-intuitive anymore.

7. Social and classical afterlife

But there also seems to be another possible reaction to a conventional objection, which is however available only within a specific metaphysical framework. It seems to be clear that if there is an almighty creator who enables you to carry your life on in a community of others, including those who passed away before you, neither afterlife nor personhood seems to be a matter of convention. I believe that it is a big advantage of the PLV theory that it is suitable with both physicalist/non-classical-afterlife and theistic/classical-afterlife metaphysics and shows also the difference of possibilities of an afterlife in each of them, where on one hand there is a limited afterlife without survival on the side of the physicalist/no-classical-afterlife metaphysics and a full afterlife on the side of theistic/classical-afterlife metaphysics. This reflects also my analysis of the benefits of afterlife belief. I showed there in which way there are premium psychological benefits in believing in classical afterlife, which nevertheless demand sacrificing beliefs preventing us from embracing theistic worldview (which is for many of too great value).

8. PLV as reflectively equilibrated

The rather conventional character of person is probably the highest price one has to pay for the possibility of embracing the non-conventional

afterlife. Schechtman shows that there are also beliefs of a more metaphysical nature that one has to sacrifice to adopt PLV. The most important one is probably that organism is not an "object of everyday life", but more some theoretical abstraction from the totality of our experience (Schechtman 2014, 183-186). This seems to be a quite counterintuitive statement and Schechtman makes lot of effort to significantly reduce its counterintuitiveness. Her argument is in this point quite complex and therefore unfortunately unsuitable to present it here.

Arguably, it is hard to compare values of various beliefs and I don't have the illusion that everyone is ready to make the trade-off suggested above. But on the other hand, I suggest that the price is reasonable enough to consider it as a real option and that PLV could be included in the list of theories of personal identity which fit the criterion of reflected equilibrium best.

9. Conclusion and future direction

In my paper, I discussed the topic of afterlife i-concerns, meaning identity-presupposing concerns that go beyond the continuity of our biological body. I showed that there are indeed such concerns and that we have practical and ethical reasons to maintain them. Then I presented the person-life view as the best candidate to implement such ideas and show how it works differently within the different ontological frameworks. I tried to show that the final position that the person-life carries on after death seems to be reasonable from the perspective of reflective equilibrium. Nevertheless, there are still a lot of open questions, which could invalidate this conclusion. It could turn out that there is after all a possibility to rationally defend one's i-concerns even when there is no link between them and a concept of a person. One could potentially interpret our practices which I linked to afterlife i-concerns without invoking them. There is of course also a possibility that PLV implies still more counterintuitive beliefs that Schechtman is not aware of. Lastly, another theory could be developed which is more intuitive than PLV and could adopt a non-conventional afterlife at the same time. I hope that this paper encourages more vivid debate about such issues.

Acknowledgements

The content of this paper is partly based on my previous conference paper 'Afterlife from a practical point of view' delivered at Y*oung Philosophy Graduate Conference 2017: The Character of Current Philosophy and Its Methods*, 2017. I am greatly indebted to two anonymous reviewers for their helpful comments on earlier versions of this article.

References

BĚLOHRAD, R. (2014): Reactions and Debate I: On Schechtman's Person Life View. *Ethical Perspectives* 21(4), 565-579.

BĚLOHRAD, R. (2016): *Lidské identity, lidské hodnoty* [Human identities, human values]. Prague: Dybbuk.

CHIONG, W. (2005): Brain Death without Definitions. *Hastings Center Report* 35(6), 20-30.

GREEN, J. (undated): Resurrection. In: *Internet Encyclopedia of Philosophy*. URL: http://www.ieutm.edu/resurrec/. [Accessed on 05/05/2017.]

HASKER, W. & TALIAFERRO, C. (2014): Afterlife. In: Zalta, E. N. (ed.): *The Stanford Encyclopedia of Philosophy*. URL: https://plato.stanford.edu/archives/win2014/entries/afterlife/. [Accessed on 05/05/2017.]

JAMES, P. D. (1992): *The Children of Men*. London: Faber and Faber.

JOHNSTON, M. (2010): *Surviving Death*. Princeton: Princeton University Press.

KORSGAARD, C. M. (1989): Personal Identity and the Unity of Agency: A Kantian Response to Parfit. *Philosophy & Public Affairs*, 101-132.

LEWIS, D. (1998): Possible Worlds. In: Laurence, S. & Macdonald, C. (eds.): *Contemporary Readings in the Foundations of Metaphysics*. Oxford: Blackwell Publishers.

MACKENZIE, C. & ATKINS, K. (eds.) (2008): *Practical Identity and Narrative Agency*. New York: Routledge.

ROBSON, S. & WHEATESTONE, R. (2017): British Chef Who Died in Syria 'Killed Himself to Stop ISIS Taking Him Prisoner'. *Mirror*. URL: http://www.mirror.co.uk/news/uk-news/british-chef-killed-syria-fighting-9733350. [Accessed on 06/04/2017.]

SCHECHTMAN, M. (1997): *The Constitution of Selves.* Ithaca: Cornell University Press.

SCHECHTMAN, M. (2014): *Staying Alive: Personal Identity, Practical Concerns, and the Unity of a Life*. Oxford: Oxford University Press.

SCHEFFLER, S. (2013): *Death and the Afterlife*. Edited by N. Kolodny. Oxford: Oxford University Press.

SHOEMAKER, D. W. (2007): Personal Identity and Practical Concerns. *Mind* 116(462), 317-357.

TETENS, H. (2015): *Gott denken. Ein Versuch über rationale Theologie* [*To think God. An Attempt at a Rational Theology*]. Ditzingen: Philipp Reclam jun. Verlag.

TETENS, H. & SCHOLL, J. (2016): *Der Philosoph Holm Tetens (Musik und Fragen zur Person).* Deutschland funk. URL: http://ondemand-mp3.dradio.de/file/dradio/2016/12/25/zwischentoene_mit_holm_tetens_vom_25122016_mit_musik_dlf_20161225_1330_4fd25a38.mp3. [Accessed on 09/05/2017.]

ZIMMERMAN, D. (2013): Personal Identity and the Survival of Death. In: Bradley, B., Feldman, F. & Johansson, J. (eds.): *The Oxford Handbook of Philosophy of Death.* Oxford: Oxford University Press.

Organon F 25 (2) 2018: 215-243

Self-organization, Autopoiesis, Free-energy Principle and Autonomy

TEODOR NEGRU[1]

ABSTRACT: The aim of this paper is to extend the discussion on the free-energy principle (FEP), from the predictive coding theory, which is an explanatory theory of the brain, to the problem of autonomy of self-organizing living systems. From the point of view of self-organization of living systems, FEP implies that biological organisms, due to the systemic coupling with the world, are characterized by an ongoing flow of exchanging information and energy with the environment, which has to be controlled in order to maintain the integrity of the organism. In terms of dynamical system theory, this means that living systems have a dynamic state space, which can be configured by the way they control the free-energy. In the process of controlling their free-energy and modeling of the state space, an important role is played by the anticipatory structures of the organisms, which would reduce the external surprises and adjust the behavior of the organism by anticipating the changes in the environment. In this way, in the dynamic state space of a living system new behavioral patterns emerge enabling new degrees of freedom at the level of the whole. Thus, my aim in this article is to explain how FEP, as a principle of self-organization of living system, contributes to the configuring of the state space of an organism and the emergence of new degrees of freedom, both important in the process of gaining and maintaining the autonomy of a living organism.

KEYWORDS: free-energy principle – self-organisation – autonomy – autopoiesis.

[1] Received: 19 February 2018 / Accepted: 11 April 2018

✉ Teodor Negru

Faculty of Philosophy, Theology and Religious Studies
Radboud University Nijmegen, Erasmusplein 1
6525 HT Nijmegen, The Netherlands

e-mail: theonegru@yahoo.com

In the current literature, the free-energy principle (FEP) is approached in the context of predictive coding theory, which provides an explanatory framework about how the brain works, understood as "an inference machine that actively predicts and explains its information" (Friston 2010, 129). This means that the brain does not passively receive information about the world, but it develops a model of the surrounding world, which it permanently adjusts based on the information received from the environment. According to this theory, in order to minimize surprises, the brain makes predictions about what will happen – or indeed, is happening at the moment. An important role in this process is played by the perception-action dynamics, which actively contributes to predicting the changes in the reality: thus, perception optimizes predictions by inferring the hidden causes of the external changes whereas by action the error of the predictions is minimized.

Minimizing surprises involves limiting free-energy, which is a characteristic not only of the brain but of all self-organizing systems (Friston 2009; 2010). Free-energy is an important aspect of all biological systems, because, from the thermodynamic point of view, it is the working energy of the organism. However, free-energy can be understood from the information theory perspective, as a function of both sensory and internal states of organism. In this context, minimization of free-energy involves increasing the probabilistic information relating to the system's exchanges with its environment and the external causes of those exchanges. In other words, free-energy is considered as the upper bound of surprise (Friston 2009; 2010).

Starting from here, one can say that FEP is an important aspect of the functioning process of any self-organizing system, which, in order to maintain the internal equilibrium, it needs to control the entropy resulting from the flow of information and energy exchanges with the world. FEP is considered a consequence of the propensity of any self-organizing adaptive system to resist disorder and to maintain its identity and unity considering the external perturbations. The integrity of living systems is maintained (or is indeed defined) by placing an upper limit on the free energy of the system. This can be achieved in one of two ways; namely by changing sensory samples of the environment (i.e. sensory input) by action or by changing the internal states of the system that enabled sensory exchange to be predicted (Friston 2010). Limiting the free-energy of

a living system is a prerequisite of the survival of an organism, involving the development of some mechanisms that would anticipate the changes in the environment and reduce the surprises from the external milieu.

In this context, the goal of my paper is to debate the relevance of the free-energy principle to the problem of biological autonomy, extending the discussion from the Bayesian approach of the brain to the process of self-organization of living systems. Considering this task, the paper is divided in four parts: the first part is an overview of the principles entailed by self-organisation in the case of living systems. In this way, a comprehensive approach of what a self-organizing living system means is achieved, taking into consideration different aspects of self-organization. In the second part of the article, the process of autopoiesis, considered in some enactivist theory as the origin of life, is approached from the point of view of the self-organisation principles, considering autopoiesis as a minimal case of self-organisation. Further, in the third part, the discussion about autopoiesis and self-organization is completed by discussing how FEP is involved both in the emergence of autopoietic systems and, in general, in the self-organization of any living system. Starting from here, in the last part, I discuss the role of FEP in gaining autonomy of a living system, considering two aspects. On one hand, I approach the way FEP contributes both to the internal self-organization of a living system, which in the autopoietic tradition is known as organizational closure, and to the emergence of its degrees of freedom, considering that any organism is also a dynamical system. On the other hand, the autonomy of a living system will be approached taking into account that any system has a boundary, which, in the case of a living dynamical system is a Markov blanket that provides a peculiar type of coupling of the organism with the world. Thus, my aim is to show that the issue of autonomy of the autopoietic theory can be completed by its approach from the perspective of FEP and dynamical system theory. In this way, a new account of autonomy of living systems is proposed, which takes into consideration not only the recent findings of autopoietic tradition, such as organizational theory, but also the research from dynamical system approach of living systems.

1. Self-organisation in living systems

At the origin of life lies self-organization of living matter, which entails the aggregation of molecules in a coherent structure, which would resist perturbations from the environment. According to the current research self-organization is a ubiquitous process, which can be found all over in nature both in inanimate forms, and in the realm of living system. For instance, self-organizing structures can be dissipative, such as hurricanes or dust devils that emerge in certain circumstances and last as long as certain conditions are met (Juarrero 2010b, 257). But self-organizing systems can also be flexible structures with the ability to evolve and self-maintain (Barandiaran & Moreno 2008, 327). In this case, the maintenance of the system is achieved by adapting the internal behavior to the changes in the environment and influencing the external conditions. These are the living systems, which, as self-organizing systems, involve a set of principles that are interdependent and operate spontaneously.[2] Together, these principles contribute to the emergence of an autonomous living system.

1.1. Principle of systemicity

The result of self-organization is the emergence of a system, meaning the configuration of some relatively stable structural assemblies, with a unitary behavior. Such a system is characterized by multistability (Camazine 2003, 34) or metastability (Nicolis & Prigogine 1977, 462), which entails the existence of several steady states the system can have, depending on the external conditions and parameters influencing the system. Thus, self-organizing living systems are not rigid structures but they involve a certain flexibility that allows for their fluctuation between certain states (Juarrero 1999, 111).

Consequently, a self-organizing system is a combination of stability and instability. This means that it is a structure, which, on one hand, obeys the deterministic laws of classical physics, exhibiting predictable behaviors,

[2] In other words, the difference between dissipative and biological systems is that in the former case, self-organization is maintained by the energy flow from outside, whereas in the latter, self-organization comes from inside the organism as a consequence of its internal organization (Ruiz-Mirazo & Moreno 2004, 238).

and, on the other hand, it is considered statistically unstable, enabling the emergence of new behaviors (Pattee 1988, 328).

Moreover, a self-organizing system involves the fact that certain elements are configured in a structure in which each part has a certain function it would exercise in order to maintain the whole.[3] This means that the elements of the system are selected, in order to be part of the new whole according to the powers they are assigned. Hence, exercising the powers of the parts depends on the functioning of the whole as well as on how they contribute to the integrity of the system. Just as, for instance, certain organs or functions of the living systems are enhanced, whereas other are diminished, according to the contribution to the survival of the organism.

Last but not least, systemicity involves the emergence of some forms of unity and identity of the system. The functioning of a whole involves the unity of processes and its actions. Unity is a consequence of the coherence of the system functions that converge towards the achievement of the same purpose, which is its survival. Identity is a consequence of the fact that processes and actions belong to the same whole. Both the unity and identity of the organism are operational, as they are the result of the internal processes of the system, which contribute to maintain its integrity.

1.2. Principle of spontaneity

Living matter has the property to self-assembly in organized structures, which would resist to the entropy of the surrounding world. An important characteristic of the self-organization of the living matter is the spontaneity of the elements coupling, which is carried out without the contribution of an external force or an internal generating principle. In other words:

---

[3] The part-whole relation can also be approached from the perspective of their properties. Thus, the system can be seen as the total amount of the properties of its parts: "A system is a group of entities with some collective property (...) Maintaining the system is thus maintaining the collective property" (Newton 2000, 92). To put it differently, between the properties of the components and those of the system there is a relation of dependence. This means that "in a system, (...), the properties of the components depend on the systemic context within which the components are located" (Juarrero 1999, 109).

> Self-organizing systems, (…), have need for neither homunculus-like agents located inside a complex system nor any kind of cosmic instruction from the outside ordering the parts around, telling them what to do and when to do it. (Kelso & Engstrøm 2006, 93)

This means that self-organizing systems do not need the existence of a self (Kelso 1997, 8), a program (Thelen & Smith 1998, 281) or an external cause that would conduct the coupling of elements. In the self-organizing process, the coupling of elements is carried out spontaneously, without a control center conducting this process. And the laws under which coupling elements is carried out result from the very process of arranging the elements.

Moreover, in the case of living systems, spontaneity is a characteristic of the responses of the organism to the environmental challenges. Behavioral patterns emerge spontaneously without the mediation of a centralizing cognitive structure such as consciousness, which would generate a conscious mediated response to the environmental changes. In other words, living organisms have the ability to spontaneously self-organize under the pressure of environmental constraints, which determine the configuration of the state space of the organism and emergence of a behavioral response.

1.3. Principle of non-linearity

Self-organization enables the emergence, at the level of the whole, of some properties that the independent parts do not have. This means that the whole is not a mere addition of its parts. The aggregation of the elements determines the emergence of some new functions and powers, in the system, which do not represent the mere addition of the characteristics and powers of elements.[4] Aggregation of the elements in a coherent configuration enables the emergence of a higher-order organization of the whole, which exhibit a state space with a high-order dynamics than of the component states. This means that the whole has degrees of freedom greater than

[4] In other words, "…dynamical processes provide empirical evidence that wholes can be more than just epiphenomenal aggregates reducible to the sum of their component parts. The newly organized arrangement shows emergent macroscopic characteristics that cannot be derived from the laws and theories pertaining to the microphysical level" (Juarrero 2010, 257).

those of its components. That is to say, it has alternatives of action and response to environmental challenges, more complex than the sum of alternatives of response of its parts considered independently.

The emergence of the new properties is a consequence of the non-linearity which is a characteristic of the biological world.[5] Non-linearity implies the unpredictability of the changes within the system, which means the emergence of new effects that cannot be deduced from the characteristics of the parts. This is possible because, in the self-organizing process, qualitative shifts emerge at the level of the whole that enable the enlargement of its state space and access to some new states by the system as a whole.

To put it differently, in the phase shifts of the self-organizing process "similar causes can have different effects and different causes similar effects; small changes of causes can have large effects, whereas large changes can also result in only small effects" (Fuchs 2007, 853). These shifts that determine new levels of self-organization to emerge are the consequence of the control parameters, which exceed some critical values under the action of the aggregate variables of the system. This determines the shift of the organizing patterns of the system and, implicitly, of the dynamics of its basic components meaning the emergence of new patterns of action.

1.4. Principle of circular causality

Self-organization consists not only in the aggregation of some components but it also involves modeling the dynamics of these components by the new emerging whole. Thus, the parts and the whole are in a mutually conditioning relation, which entails that the parts constitute the whole and in turn are constrained to adopt certain behavior by the whole. This circular causation relation determines the emergence of the micro-dynamics of components from the macro-dynamics of system, which in turn will determine the micro-level dynamics. In Kelso and Engstrøm's description (2006, 114-115), the circular causality relation involves the coordination of three levels: the "lower level" of the components interaction that results

[5] According to Thelen and Smith (1994, 58): „Self-organization is not magic; it occurs because of the inherent nonlinearities in nearly our physical and biological universe."

from macro-level (upward causation), the "upper level" that plays a boundary condition role, which constrains the dynamics of the coordinating elements (downward causation), and the "middle level" made of the coordinating patterns between the macro- and the micro level. Thus, circular causality is approached from the point of view of the dynamics between the upward and downward causation.

Approached from the perspective of the patterns created by the system, the circular causality relation entails modeling the dynamics of basic level by the patterns of action it creates at the higher level. In other words, from the coordination of the basic components a pattern of action results that integrates all parts of the system in a whole, which share a common dynamics. Thus, the coordination of the components of the system by its macro-patterns enslaves the behavior of the parts achieving the behavior of the whole. From this perspective, the circular causality is based on the slaving principle (Haken 1983), according to which the formation of the slowest microscopic patterns resulting from the fastest dynamics at the microscopic level involves decreasing the degrees of freedom of the system components and reducing the states of the system to only a few.

The reciprocal causation among the levels of a complex system can be understood in terms of the coupling or dynamics between microscopic (fast) and microscopic (slow) order parameters. This means that the microscopic fluctuations of the system that constitute its behaviour determine the emergence of macroscopic order parameters – that enslave the microscopic degrees of freedom (Bruineberg & Rietveld 2014, 5). This synergetic, enslaving principle rests upon circular causality and organises a system's degrees of freedom onto a low dimensional manifold that contains the macroscopic order parameters.

1.5. Principle of adaptivity

Self-organization entails the emergence of a living system that is fit to the condition of the environment within which it lives. Thus, in the self-organization process are involved both the internal parameters of the system, upon which the internal coherence of the system processes depend, and the external ones, a consequence of the environmental conditions. The external parameters determine the selection of those functions and powers that enable the whole to adapt to the changes and fluctuation to the

environment. Depending on how it adjusts to the environment, the system is also characterized by certain robustness, which is "the system capacity to maintain its organization in the face of internal and external perturbations" (Barandiaran & Moreno 2008, 331). From this perspective, self-organization implies the emergence and selection of those patterns that would provide the robustness of the organism under the circumstances of environmental changes.

An important role in the process of adaptivity is the way the system is linked to its milieu. As living systems have emerged and developed for generations within a certain milieu, they are coupled initially and structurally with this milieu. Structural coupling involves that the organism, by means of its organs, perceives directly the changes in the environment and is prepared to provide optimal response to such changes. Thus, a living system is not an isolated structure within the environment it lives, but "the external structure or boundary conditions of complex systems are as much as part of the complex system as the internal structure" (Juarrero 2010a). This is what enables the living system to interact with the milieu, not only passively, by receiving information from the outside, but actively as well, by transforming the milieu where it lives. In other words, structural coupling involves symmetry between the system and the world, meaning their mutual influence (Di Paolo 2010, 50). By this mutual influence, the organism acquires the information necessary to preserve a state of dynamic equilibrium with the world. Consequently, adaptability involves regulation of an organism according to an interactive cycle between the living system and the world (Barandiaran & Moreno 2008, 335).

Structural coupling is facilitated by the emergence, in the process of self-organization, of a boundary between the living system and the world. This boundary (Ruiz-Mirazo & Moreno 2004, 244-245) is a demarcation between the system and the world, and it also enables exchanging information between the organism and the world. Boundary delimitates the internal space of the organism, its inner vital field whereby the system gains the autonomy of its internal processes. Moreover, boundary is endowed with receptors sensitive to the changes in the environment and with structures that enable exchanges with the external milieu, which would facilitate the adaptation of the organism.

1.6. Principle of optimality

Self-organization involves not only the emergence of simple responses of living systems to the environmental challenges. It also involves selecting those responses that are the most appropriate to the challenges in the milieu. In other words, the patterns of action emerging as a result of self-organization are the most efficient to answer to the changes in the environment (Bruineberg & Rietveld 2014, 5, note). This means that, on one hand, the behavioral patterns are generated according to the energetic possibilities of the system. That implies that the organism has the resources required to configure and complete the pattern of action. On the other hand, the patterns generated in the state space of the system should respond to as many parameters as possible of those, which influence the system. This means that state space of a system should also be made of optimal states to be occupied by the system in order to provide optimal responses. Thus, the survival of the organism means generating the optimal patterns, according to the energetic abilities of the organism, which would enable the coverage of as many variants as possible to respond.

1.7. Principle of thermodynamic non-equilibrium

The propensity of self-organizing systems is to maintain a state of stability, being from a thermodynamically point of view in a state of non-equilibrium due to the energy and information flows they are subjected to. Stability does not require the system to be in absolute rest, as this would mean the end of the system activity.[6] Stability is merely a transitory state, until the perturbation of the system variables and configuration of another stable state. This means that, "In self-organisation, the system *selects* or is *attracted* to one preferred configuration out of many possible states…" (Thelen & Smith 1994, 57). It results that, in case of living systems that evolve in time, self-organization involves reaching a dynamic (nonequilibrium) steady-state, considering the environmental conditions and the degree of development of the organism at that time.[7]

---

[6] In Kauffman's terms, this means that "There is no agency at equilibrium" (Kauffman 2000, 66).

[7] In physics, the sort of stability associated with self-organisation and autopoiesis is referred to as 'non-equilibrium steady-state'. In other words, an ergodic or invariant

To put it differently, from the perspective of energetic and information exchanges, a living system is an open system which is in an ongoing flow of exchanges with the environment. This is due to the structural coupling that determines the ongoing interaction and permanent exchanges with the environment. As the system receives continuously energy from the exterior, it can maintain its current state, but, at the same time, its internal organization is in danger. This happens because, if a too large quantity of energy enters the system, the system entropy increases until the extinction of the system. Therefore, the problem a living system faces is how to maintain low entropy within the system and to control the energy and information flow to which it is subjected (Ruiz-Mirazo & Moreno 2000, 212-213).

The control of the flow exchanges with the exterior involves that the organism reaches a homeostasis state, whereby it gains dynamic equilibrium with the environment (Newton 2000, 93). In terms of dynamical systems, homeostasis means controlling the internal variables of the living system and maintaining them within some boundaries so that the system has a constant behavior oriented towards its preservation.

* * *

To conclude, self-organization of living system implies spontaneous emergence of a systemic whole with operational unity and identity, which are given by the coherent functioning of its internal processes. The properties of this systemic whole are more complex than those of its parts and cannot be reduced to the properties of the components. The newly emerged whole is characterized by a state of dynamic stability. This is the consequence of the internal dynamics of the system, which is given by the reciprocal causation relation between the parts and the whole, and of the external dynamics, i.e. the state of thermodynamic non-equilibrium between the organism and the environment. Last but not least, a self-organizing living system is a system adapted to the environment, enabling multiple and complex behavioral patterns that are the most appropriate to responding to the

---

property that is far from equilibrium and entails a succession of stable states. In what follows, we will also refer to this nonequilibrium steady-state as a 'dynamic equilibrium'.

external changes. Considering all this, one can say that self-organized living systems are characterized by a dynamic, non-linear and multidimensional state space, which is configured, taking into account the adaptive skills of the organism and the external parameters.

2. Self-organisation and autopoiesis

Taking into account that self-organization is an essential process in the emergence and maintenance of life, an important issue for understanding how living organisms function is the relation with the process of autopoiesis, considered to have a significant role in the emergence of life. According to Maturana (1987), the two concepts have nothing in common, that is he would "never use the notion of self-organization […]. Operationally it is impossible. That is, if the organization of a thing changes, the thing changes." This means that self-organization involves more than a re-organization within the system, but it involves a complete change of the system. In the terms of Collier (2004, 168), who analyzes the relation between the two concepts, autopoietic systems are able of self-governing and re-arranging their parts but cannot produce a new organization. In addition, Collier (2004, 151) shows that, according to Maturana and Varela (1980), the process of autopoiesis implies the existence of an organized self, whereas self-organization can be achieved in the absence of such a self. Notwithstanding, a closer analysis of Maturana and Varela's theory (1980), from the perspective of self-organization principles, shows the complementarity of the concepts of autopoiesis and self-organization.

According to the classical autopoietic theory developed by Maturana and Varela (1980, 79-80), a living system is an autopoietic machine, which has the capacity to maintain its internal variables constant. This means that living organisms are homeostatic systems that maintain their internal organization invariable. Thus, what differentiates autopoietic systems from other systems is the capacity to self-produce, which means the capacity to maintain their organization by themselves. This is possible because the internal organization of such system is a network of processes that generates and maintains the internal components, which contribute to the functioning of such processes. Hence, the internal processes of the

system form an interconnected network that also generates the boundary of the system, which gives unity to the system.

Starting from here, autopoiesis is regarded as a "specific instance" (Varela 1992, 6) of self-organization, that is to say, a type of self-organization characterizing minimal living systems. As a self-organizing process, autopoiesis constitutes the identity of the system: thus, the identity of an autopoietic system is the result of invariant patterns emerging within the system due to its internal organization. These invariant patterns provide stability and continuity to the system, despite the energy flows that continually affect the living system.[8]

Moreover, as a self-organizing constitutive process, autopoiesis is characterized by the dynamics between the local component and global whole, meaning by reciprocal causation between "the local rules of interactions (…) and the global properties of the entity" (Varela 1992, 6).[9] Reciprocal causation is a circular causality where the components interaction determines the production of the whole, which, in turn, determines the maintenance of the components.

Furthermore, another basic characteristic of an autopoietic system is that as biological system it should have a certain relation with the environment. This relation is defined as a reciprocal coupling (Varela 1992, 7), whereby the system, on one hand, separates from the environment in order not to become one with it, and, on the other hand, maintains energy and information exchanges with its external milieu.

Last but not least, one can add that an autopoietic system is not the result of some external force that would create it, nor is it an internal homunculus, it does not lay at the basis of its organization. Even if any living system involves a self – which in its minimal form looks like a coherent pattern

---

[8] In terms of the dynamical systems theory, this means that attractors of a system are autopoietic or self-creating, the attractors being the consequence of the system propensity to minimize its entropy (Friston & Ao 2011, 7).

[9] Starting from here, which is from the perspective of the process of autopoiesis, self-organization can mean "(a) local-to-global determination, such that the emergent process has its global identity constituted and constrained as result of local interactions, and (b) global-to-local determination, whereby the global identity and its ongoing contextual interaction constrain the local interaction" (Froese & Ziemke 2009, 497). In other words, the process of autopoiesis can be described, in dynamical systems terms, as the result of the dynamics between downward and upward causation.

emerging from the interaction of local components – this is the result of its internal organization (Varela 1992, 11). The internal organization of a living system emerges spontaneously taking into account only the coherence of the processes of the system and the circumstances in the environment.

Notwithstanding, in the later approaches of autopoietic theory, an important characteristic of a living system, which distinguishes it from other self-organizing systems, is self-determination (Moreno & Mossio 2015; Mossio & Bich 2017). According to this approach, biological organisms have the capacity to establish their own condition of existence, due to the circularity, which constitutes its internal organization. This means that "the organization produces effect (e.g., the rhythmic contractions of the heart) which, in turn, contribute to maintain the organization (e.g., the cardiac contractions enable blood circulation and, thereby, the maintenance of the organization)" (Mossio & Bich 2014, 1090).

Self-determination is a consequence of the closure of the organism, which has the capacity to self-constrain. In other words, the network of recursive and interactive processes that constitute the autopoiesis process is at the origin of what Varela (1979, 58) called organizational closure. Organizational closure implies that the system has the capacity to self-produce the constraints upon which its condition of existence depends (Bich 2016, 207). Approached from the perspective of the constraints generated by the internal organization of any biological system, organizational closure is understood as biological closure (Moreno & Mossio 2015, 5). Biological closure involves the fact that a biological system operates by means of the constraints it generates upon the thermodynamic flow it undergoes as open system that operates in far from equilibrium conditions (Moreno & Mossio 2015, 6). Due to biological closure, biological organisms have the capacity to self-constrain, namely to act upon their boundary conditions, which involves self-maintenance and self-determination.

To conclude, according to organizational view, self-determination is a characteristic of biological systems, which is not present in case of other self-organizing systems such as dissipative systems. This happens because: Dissipative structures posses a low internal complexity, which is precisely what enables them to *spontaneously* self-organise when adequate boundary conditions are met. In contrast to biological organisms, self-organizing systems are systems that are simple enough to appear spontaneously. (Mossio & Bich 2014, 1108)

The conclusion resulting from this is that the dissipative structures are guided by a single macroscopic constraint, being highly dependent on external conditions. Whereas biological organisms, as systems with a higher-order complexity, have the capacity to self-determine and self-maintain, due to the large number of constraints generated, which are in a close interdependence (that is they form a closure of constraints) (Moreno & Mossio 2015, 16).

From the point of view of dynamical systems, dissipative structures are considered structures dependent on external conditions (Juarrero 2015, 4). However, one may add, these are systems characterized by a limited state space, with finite and lower dimensionality. Due to this state space, they can configure only a limited number of simple patterns, as a response to the pressure of the environment. Unlike these systems, biological organisms, due to their complexity have a state space with a higher-order dimensionality, configured by the multitude of their variables. Such a state space enables the emergence of behavioral patterns with complex, and sometimes, unpredictable trajectories.

However, in both cases, the emergence of new properties of the systems, namely its nonlinearity, is due to the constraints acting onto the system. From this perspective, Juarrero (1999; 2010b) distinguishes between the context-free constraints, which are imposed from outside the system and does not generate novelty and complexity, and context-sensitive constraints, which operate as enabling constraints, determining the emergence of new properties. Context-sensitive constraints act based on the circularity relation between the part and the whole, acting bottom-up (as first-order contextual constraints), by correlating the parts of the system and enlarging its state space, and top-down (as second-order contextual constraints), by its new dynamics which the whole share with its parts. Hence, self-organization of complex systems is understood as the result of the dynamics between the context free and first-order contextual constraints, which by adding and correlating the parts determines the emergence of the new properties of the system, which provide a new dynamics to the system components (Juarrero 1999, 142).[10]

[10] In other words, self-organization involves, due to the constraints it is subjected to, the emergence of at least one bifurcation within the system, which would enable a more or less complex behavior (Hooker 2013, 764).

Consequently, enabling constraints determine qualitative changes in the whole system, enlarging the system's state space (Juarrero 1999, 143). Moreover, enabling constraints can determine the modification of the system state space, so that new trajectories can emerge and it can access new states (Hooker 2013, 761). In this context, self-determination, as a characteristic of the higher-order complexity self-organizing systems, is a consequence of enabling constraints. Self-determination refers to the possibility of a living system, due to enabling constraints on the system parameters, to generate new behavioral patterns and to configure a dynamic state space with new degrees of freedom. Organizational closure of a living system is the result of enabling constraints, which determine qualitative changes within the system.

To sum up, autopoiesis is a case of basic self-organization in the biological world, which involves all principles of self-organization. Notwithstanding, self-organization in the case of biological organisms involves mechanisms other than those in other self-organizing systems (i.e., dissipative systems). Biological organisms are self-organizing systems that are capable of self-determination due to enabling constraints. Thus, organizational closure of the living system, due to the enabling constraints of the system, exhibits a multidimensional state space, which allows the emergence of some complex behavioral patterns. Consequently, if self-organizing dissipative systems have an invariable state space, self-organizing living systems have a dynamic state space, which can be extended depending on the adaptive needs of the system.

3. The free-energy principle, self-organization and autopoiesis

One of the consequences of the self-organization of living matter in not only the emergence of a system with a coherent structure, but also with the capacity to resist ongoing perturbations from the environment. Starting from this, one can say that the FEP is an important aspect of any self-organising process (of a living system), which, as an open system, should control the energy and information exchanges with the exterior in order to not increase the system entropy. This means that without FEP living systems would not be able to exist because "the entropy of their sensory states

would not be bound and would increase indefinitely" (Friston 2013a, 2), which would result in the extinction of organisms.

Minimizing free-energy has an important role in the organism adaptivity to the environment as well (Bruineberg, Kiverstein & Rietveld 2016, 2; Kirchhoff 2016, 4). In order to survive, any living organism aims at integrating in the environment where it lives. From the dynamical system point of view, this means that from the interaction between the organism and world results a whole as an organism-environment system (Menary 2007, 42). Thus, adaptivity involves the capacity of living organisms to create a system with the world. This means that in structural coupling of the organism with the world, which implies their mutual conditioning (Di Paolo 2005), an organism-world assembly results with a common dynamics. Thus, the organism does not act as an isolated entity, which receives passively information about the environment, but it becomes a part of the world coordinating its actions with the changes in the environment.

An important role in this process of adaptation is played by the internal structures of the living system, which detect and anticipate the changes in the world. Adaptivity involves attunement of the internal processes and actions of the organism with the changes in its econiche. This means that the organism does not develop a representational model of the world based on which it acts. But the organism is itself a model of the world where it lives, having a direct relation with it (Friston 2013b, 213). This involves, on one hand, that it is endowed with skills that complement its econiche, and on the other hand, that between the internal dynamics of the organism and the external one of the environment there is a state of equilibrium or optimal grip (Bruineberg, Kiverstein & Rietveld 2016; Bruineberg & Rietveld 2014). Thus, embodied skills of organisms, a consequence of their internal organization, achieve the integration of the organism in the environment and the creation of a system with a shared dynamics with external milieu.

The systemic coupling involving that every self-organizing living system to embody an optimal model of its niche (Friston 2011), makes the organism to exhibit the best patterns of response to the external challenges (according to a variational principle of optimality). Moreover, it results from the systemic coupling of the organism with the world that the skills of the organism are directed not only towards maintaining internal organization, but also towards anticipating the changes in the environment. Thus, the organism minimizes the external surprises that may affect the system,

maintaining its activity within the boundaries of a low number of states that could ensure the survival of the system (Bruineberg, Kiverstein & Rietveld 2016, 2; Friston 2011). It results that the self-organizing living systems have the ability to change the configuration of their state space, controlling the states, which the organism can access by limiting its free-energy. This means that while functioning, the living systems aim at minimizing the surprises of entering in a certain state (Kirchhoff 2016, 4), reducing the degree of freedom of the system and its state space, by regulation of its free-energy.

It results from here that regulation, as a process that contributes to the organism adaptivity (Di Paolo 2005, 430), being a form of adaptive control (Mossio & Moreno 2010, 285), is one of the characteristics of a self-organized living system. According to the organizational theory (Moreno & Mossio 2015, 33), the mechanism underlying the regulation of living systems is explained by second-order constraints, which are different from constitutive constraints, which ensures maintenance of the organism under stable conditions. Second-order constraints emerge when the organization of the system is endangered, having the role to re-establish the internal closure of the organism. In this case, regulation involves modulation of the constitutive regime until the recovery of the closure of the organism. In this approach, regulation takes the form of a circular organization of organism: constitutive constraints are those that are at the basis of second-order constraints, and regulatory constraints by establishing a second-order closure contributes to maintaining the constitutive constraints. Thus, regulation involves decoupling from the constitutive level and increasing the complexity of organism, by means of the emergence of some new levels within the system, with new degrees of freedom.

The circular causality supported by the organism constraints is also at the basis of the mechanism of limiting its free-energy. Thus, at the level of constitutive regime, constraints that are at the basis of organizational closure harness the flow of energy of organism in order to maintain its organization, and, at the same time supports this flow (Bich, Mossio, Ruiz-Mirazo & Moreno 2015, 8). If the constitutive constraints cannot harness the free-energy of an open system, the result is the increase of its entropy. In this case, the regulatory constraints, which operate on the constitutive regime, emerge re-establishing the equilibrium within the living system.

Starting from these assumptions, one can say that FEP can be also understood as an important aspect of the functioning mechanism of the autopoietic living systems. According to Kirchhoff (2016, 3), the difference between autopoietic theory and FEP, is that the former refers to self-production and the latter refers to self-preservation. This means that from the autopoietic perspective, self-maintaining of a system is merely an internal issue, which consists in the self-production of its internal components, with no connection with its exterior. Whereas, from the point of view of FEP, self-maintaining of a living system should consider the environment within which it lives. In other words, from the perspective of autopoietic theory, self-organization of a system relates only to its internal organization, which involves maintaining internal processes and components. Furthermore, from the perspective of FEP, self-organization of a living system involves attunement of the system and world, in order to maintain the integrity of the organism, by developing a model of the world by the living system and anticipating the changes in the external milieu.

However, autopoietic theory and FEP are understood as being convergent to the extent that both have as a result maintaining a state of homeostasis of the organism (Kirchhoff 2016, 8). According to this point of view, the process of autopoiesis involves minimizing its free-energy by minimally self-produce the components of the organism so that it maintains a model of the world. Thus, organism, both by its internal processes and its actions tends to maintain structurally and functionally integrity of itself (Friston 2013a, 5).

Nonetheless, even if maintaining the internal equilibrium, despite the changes in the environment, represents a defining feature of the self-organizing biological systems (Friston 2010, 127), whereby they distinguish themselves from other self-organizing systems, introducing FEP involves that between organism and the world there is a state of dynamic equilibrium. To put it differently, homeostasis is the tendency of the organism to maintain the internal variables constant. But the steady state of an organism is not constant. It undergoes ongoing changes that imply maintaining equilibrium when moving from one state to another, depending on the quantity of free-energy from the system. Homeostasis is a state of equilibrium characteristic to simple systems that cannot access very many states and whose behavioral patterns aim at returning to the initial state. However, living organisms have a dynamic equilibrium that implies reaching of several states

of stability along with the change of external and internal parameters as a result of the energetic changes with the exterior.

Thus, instead of homeostasis, one can speak of allostasis, which means "achieving stability through change of state" (Schulkin 2003, 21). This means that living systems are characterized by dynamic stability, which implies that the system is in equilibrium among several states and configures more trajectories to reach its states in the state space. From this perspective, the role of regulatory mechanisms is not to maintain constancy of their internal milieu, but to adjust continuously their milieu in order to survive (Sterling 2012, 5).

An important role in this dynamic of regulatory process is played by the anticipation of the changes in the environment. Thus, living organisms have developed special organs (such as the brain) that would monitor the internal and external parameters of the system so as to anticipate the changes and minimize error by adjusting their behavior according to the external changes (Sterling 2012, 7). In this process, the brain as an anticipatory organ plays the role of coordinating the internal organs and their functions in order to respond as best as possible to its predictions. Thus, living organisms achieve a predictive adaptation (Sterling 2012, 8), which involves regulating the organism by anticipating the changes in the environment.[11]

Explained from a dynamical point of view, regulation consists not only in mechanisms of constantly maintaining internal variables, but it also involves an external component. That means, minimizing free-energy of the organism, as a principle of its functioning, by anticipating the changes in the environment. Prediction of external changes has as an internal correlative the prediction by the brain of the future needs of the organism. In this way, the brain creates behavioral patterns that would adjust the internal

[11] Notwithstanding, anticipation is not a characteristic of the organisms endowed with advanced cognitive skills, such as human beings. Research in biology have shown that we can also speak of predictive behaviors in the case of bacteria (Lyon 2015) or more developed animals that do not possess language, such as rats or monkeys (Pezzulo 2008). As Keijzer (2001) said, taking into account that anticipative behavior required a new macroscopic order that would control the organism, it results that all behavior is anticipative behavior. Thus, predictive adaptation is a characteristic of living organisms whereby the aim is to obtain a dynamic equilibrium with the world.

state space of the organism depending on the changes detected in the environment. Controlling free-energy involves modeling the state space of organism, its contraction or extension, so as not to occupy those surprising (i.e., high free energy) states that would endanger its function and, at the same time, to find the best responses to environmental challenges.

In conclusion, by introducing FEP as one of the principles of self-organization of a living system, it results that biological organisms, due to the system coupling with the world, are in a state of dynamic equilibrium with its milieu. This state of dynamic equilibrium involves adjusting the behavior of the organism by anticipating the changes in the environment that will affect the states of the organism. In terms of dynamical system theory, this means that state space of a living system is characterized not only by several stable states, which it occupies alternatively, depending on the external conditions. But state space of a living system is a dynamic space which can be extended or restrained depending on the organisms predictions and how it controls its free-energy. The consequence of attunement of the internal dynamics of the organism with the external one of the environment is the emergence of a dynamic state space that is configured depending on the anticipations of the organism, by adding or restraining certain states. Moreover, in this dynamic state space, depending on the abilities of the organism, several trajectories can configured in order to reach a certain state.

4. Free-energy principle and autonomy

One of the consequences of self-organization of living matter is to develop an autonomous biological system. Autonomy is the feature of the living systems to function independently of external conditioning, by creating its own conditions of existence to survive. In terms of organizational theory, autonomy of a living system can be approached from a double perspective: from the point of view of the internal functioning of the organism (this is the constitutive dimension by which identity of the organism is made up) and from the perspective of the relation the organism has with the exterior (this is the interactive dimension which refers to the system interaction with the exterior) (Moreno & Mossio 2015, xxviii). Thus, autonomy of a living system is a twofold issue, which needs to be examined

both from the perspective of the internal dynamics of the organism and from the perspective of the external one.

According to organizational theory, constitutive autonomy is the consequence of the organizational closure, which results from generating within a living system of a new causation regime that produces and maintains the internal components of the living system (Moreno & Mossio 2015, xxvi-xxviii). Thus, between the components of a living system there is an interdependence relation whereby the constitutive elements of the system mutually condition by the emergence of a network of constraints that provides the internal functioning of the organism. Understood from this perspective, autonomy means self-determination (Moreno & Mossio 2015, 5) or self-maintenance (Moreno & Mossio 2015, 9) of the organism, which entails the capacity of a living system to replace its internal components, due to its internal organization, understood as a network of constraints that provides the regeneration of the system.

From the perspective of the internal dynamics of the system, the ability of an organism to self-maintain can be understood from the perspective of the circularity relation between the lower and higher-order level of its organization. This means that the level of the basic metabolic processes generates and supports the higher-order level of processing information, which, in turn, models the behavior of the lower level. The circularity relation between the levels of the systems also determines its dynamic organization, which involves ongoing self-organization of the components of the system according to an order pattern. From the perspective of FEP, the circularity relation contributes to reducing the system entropy, by introducing a macroscopic order to the system according to a self-organizing pattern, under the pressure of environmental conditions. Thus, the free-energy of the system is controlled by the emergence of a pattern of action that would respond to the immediate needs of the organism.

Hence, the main feature of the internal organization of a living system is not merely recursive production of its components, but also creating a more extended state space. In other words, increasing the repertoire of states encompassed in its attracting set or manifold. Autonomy of living systems does not consist only in preserving its internal organization, but it also refers to the states it can access as a result of the responses to environmental challenges that the organism provides as a whole. Thus, understanding autonomy of a living system should take into account that the state

space of a living system is a dynamical one. This means, as we have already seen, that the state space of a living system can be extended or restrained due to the anticipatory structures of the organism that can mobilize its resources in order to configure some new patterns of action. Thus, living systems have the ability to access new states and control new trajectories that encompass such states thus gaining new degrees of freedom.

In other words, by limiting free-energy a new order is introduce in the system. This means that degrees of freedom of the components are restrained, according to the new order, whereas, at the level of the whole, degrees of freedom of the system as a whole emerge. FEP contributes thus to the emergence of the degrees of freedom of the system as a whole, by creating a multidimensional state space and patterns of action whereby the system entropy is reduced.

As mentioned before, in agreement with organizational theory, autonomy of a living system is not merely an issue of internal organization, but it also depends on how the organism couples with the world. Depending on the coupling with the external world, the organism receives information from it and has the possibility to respond to the environmental challenges. An important role in the coupling of organism with the world is played by the boundary of organism. This physical border which is the result of internal processes of organism traces the boundaries between the internal space of the organism and the surrounding world, and also facilitates the communication between them (Moreno & Mossio 2015, xxvii). The circularity relation between the internal processes of a living system, which constitute its physical boundary, contributes both to the preservation of internal processes and to the constitution of the system identity (Moreno & Mossio 2015, xxvii).

From the point of view of FEP, the physical boundary of the organism has a double role: an endogenous one of controlling the internal energies of organism. And from this perspective, one can say that FEP contributes to constituting the identity of organism by controlling its internal energy and redirecting it towards the patterns of action that would provide maximum efficiency of the system actions. However, from an exogenous point of view, the boundary of the organism plays the role to control the external flow of free-energy, filtering the quantity of energy that enters the organism. Thus, FEP contributes to the unity of the living system, protecting its internal integrity.

Depending on the complexity of the organism, this physical boundary can enable the coupling of the organism with the world on several levels. An example of such boundary is the cell membrane, which is a permissible selective structure (Ruiz-Mirazo & Moreno 2004, 245) that contributes not only to setting the boundaries between the organism and the world, but also to the adaptation of the organism by detecting the changes in the environment. Similarly, the nervous system not only enable the energetic and information interaction of the living system with the world, but also a direct coupling with it, which increases the possibilities of the organism to respond to environmental challenges.

In terms of dynamical systems, the boundary separating a self-organized complex system from its milieu is called Markov blanket. A Markov blanket is defined as a set of states delimiting the internal states of a living organism from its external ones (Friston 2013a, 2). According to this description, the states that form Markov blanket are linked with the internal ones of the system, forming thus a network made of parents, children and children's parents. The internal states are a probabilistic representation of the external ones being thus able to anticipate external changes (Friston 2013a, 7) and to put the system within a certain state, which would ensure its survival. Consequently, the role of Markov blanket is to stabilize the internal states of the system and to reduce the free-energy resulting from the dynamics between the internal and external states (Friston 2013a, 4). As boundary of the system, Markov blanket represent a dynamic demarcation between the organism and the world, which enables the systematic coupling with the environment and gaining a dynamic stability by anticipating the states of the system that are to be accessed.[12]

[12] The very existence of a Markov blanket – that underwrites a separation between the system and its eco-niche – means that the internal states can be interpreted as a probabilistic representation of the external states. This representational interpretation allows one to talk about the system anticipating or predicting external changes. Mathematically, this follows from the fact that the dynamics that maintain the integrity of the Markov blanket are gradient flows on something called Bayesian model evidence (i.e., negative free energy). This means the very existence of a Markov blanket – and implicitly the system – will look as though the Markov blanket is stabilising the internal states of the system.

At the level of organism, there can be several Markov blankets (Friston 2013a, 10): cell surface, neuronal systems, etc. This means that, depending on their complexity, organisms can exhibit multiple levels of limiting free-energy. Thus, membrane can be approached as a boundary, which separates intracellular states from the extracellular ones, hidden from the internal states (Friston 2013a, 2). Communication between the two is carried out by means of sensory states (corresponding to the states of receptors and ion channels), which receives the changes within the external states, conveying internal states to them, and active states (corresponding to various transporter and cell adhesion processes), whereby internal states act upon the external states (Friston & Po 2011, 2; Friston 2013a, 2). This circular relation allows for the regulation of the integral states of single cell organism in agreement with the external changes, by configuring some behavioral patterns, made up of sensory states and active states parameters. Moreover, active states are those that bound entropy of the system, providing thus the integrity of the Markov blanket (Friston 2013a, 5). This means that state space of a living system is made up of the active states of the system, meaning of the states, which the system can access as a response to the environmental challenges.

To conclude, autonomy of a living system entails taking into account both the internal dynamics of the organism the result of circular causation of the internal parts, and the external one, between the organism and its milieu (across the Markov blanket). Minimizing the system free-energy contributes actively to gaining the autonomy of living systems by configuring and preserving the state space of the system within certain boundaries. State space of living systems is a multidimensional one enhanced by the anticipatory structures of the organism, which enable the access to new states based on predictions of environmental changes. This multidimensional state space determines the emergence of some behavioral patterns with new degrees of freedom. Thus, FEP, as principle underlying the autonomy of living systems, determines modeling the state space of organism, depending on the responses that such organism can provide and the emergence of new degrees of freedom as a result of the complexity of emerging behavioral patterns.

5. Conclusion

To conclude, FEP has an important role not only in the functioning of self-organized living systems but also in underwriting the autonomy of living systems. Minimizing free-energy is a process that contributes both to the constitution of the internal organization of the system but also to the systemic coupling of the system with the world. From the perspective of the system constitutive dimension, enabling constraints characterizing the internal organization of the system determines the emergence of a multidimensional state space, with degrees of freedom higher than those of its components. From the perspective of the interactive dimension, FEP contributes to limiting the energy entering the system by anticipating the changes in its external milieu. The coordination of internal states with external states (across the Markov blanket) is performed by behavioral patterns, which also performs attunement of the internal regulating dynamics of free-energy with the external one. Thus, the autonomy of a living system depends on its multidimensional state space and the degrees of freedom of its behavioral patterns emerging from this state space.

References

BARANDIARAN, X. & MORENO, A. (2008): Adaptivity: From Metabolism to Behavior. *Adaptive Behavior* 16(5), 325-344.

BICH, L. (2016): Systems and Organizations: Theoretical Tools, Conceptual Distinctions and Epistemological Implications. In: *Towards a Post-Bertalanffy Systemics*. Springer International Publishing, 203-209.

BRUINEBERG, J. & RIETVELD, E. (2014): Self-Organization, Free Energy Minimization, and Optimal Grip on a Field of Affordances. *Frontiers in human neuroscience* 8, 599.

BRUINEBERG, J., KIVERSTEIN, J. & RIETVELD, E. (2016): The Anticipating Brain Is Not a Scientist: The Free-Energy Principle from an Ecological-Enactive Perspective. *Synthese* 1-28.

CAMAZINE, S. (2003): *Self-Organization in Biological Systems*. Princeton University Press.

COLLIER, J. (2004): Self-Organization, Individuation and Identity. *Revue internationale de philosophie* 228, 151-172.

DI PAOLO, E. A. (2005): Autopoiesis, Adaptivity, Teleology, Agency, *Phenomenology and the Cognitive Sciences* 4(4), 429-452.

DI PAOLO, E. A. (2010): Overcoming Autopoiesis: An Enactive Detour on the Way from Life to Society. In: *Advanced Series in Management*. Emerald Group Publishing Limited, 43-68.

FRISTON, K. (2009): The Free-Energy Principle: A Rough Guide to the Brain? *Trends in Cognitive Sciences* 13(7), 293-301.

FRISTON, K. (2010): The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience* 11(2), 127-138.

FRISTON, K. (2011): Embodied Inference: Or I Think therefore I am, if I am What I Think. *The Implications of Embodiment (Cognition and Communication)*, 89-125.

FRISTON, K. (2013a): Life as We Know It. *Journal of the Royal Society Interface* 10(86), 20130475. http://doi.org/10.1098/rsif.2013.0475

FRISTON, K. (2013b): Active Inference and Free Energy. *Behavioral and Brain Sciences* 36(03), 212-213.

FRISTON, K. & AO, P. (2011): Free Energy, Value, and Attractors. *Computational and Mathematical Methods in Medicine* 2012, doi:10.1155/2012/937860.

FROESE, T. & ZIEMKE, T. (2009): Enactive Artificial Intelligence: Investigating the Systemic Organization of Life and Mind. *Artificial Intelligence* 173(3-4), 466-500.

FUCHS, C. (2007): Self-Organizing System. *Encyclopedia of Governance*. London: Sage, 863-864.

HAKEN, H. (1983): *Synergetics. An Introduction. Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry, and Biology*. 3rd Edition. Berlin: Springer.

HOOKER, C. (2013): On the Import of Constraints in Complex Dynamical Systems. *Foundations of Science* 18(4), 757-780.

JUARRERO, A. (1999): *Dynamics in Action: Intentional Behavior as a Complex System* (Vol. 31). Cambridge, MA: MIT Press.

JUARRERO, A. (2010a): Complex Dynamical Systems Theory. *Cognitive Edge Network*. www.cognitive-edge.com.

JUARRERO, A. (2010b): Intentions as Complex Dynamical Attractors. In: Aguilar, J. & Buckareff, A. (eds.): *Causing Human Actions: New Perspectives on the Causal Theory of Action*. Cambridge, MA: MIT Press.

JUARRERO, A. (2015): What Does the Closure of Context-Sensitive Constraints Mean for Determinism, Autonomy, Self-Determination, and Agency? *Progress in Biophysics and Molecular Biology* 119(3), 510-521.

KAUFFMAN, S. A. (2000): *Investigations*. Oxford: Oxford University Press.

KEIJZER, F. (2001): *Representation and Behavior*. Cambridge, MA: MIT Press.

KELSO, J. S. (1997): *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press.

KELSO, J. A. & ENGSTRØM, D. A. (2006): *The Complementary Nature*. Cambridge, MA: The MIT Press.

KIRCHHOFF, M. D. (2016): Autopoiesis, Free Energy, and the Life–Mind Continuity Thesis. *Synthese*, 1-22.

LYON, P. (2015): The Cognitive Cell: Bacterial Behavior Reconsidered. *Frontiers in Microbiology* 6, 264.

MATURANA, H. (1987): Everything Is Said by an Observer. In: Thompson, W. Ir.: *Gaia, a Way of Knowing: Political Implications of the New Biology*. Great Barrington, MA: Lindisfarne Press, 65-82.

MATURANA, H. & VARELA, F. J. (1980): *Autopoiesis and Cognition. The Realization of the Living*. Dordrecht: D. Riedel.

MENARY, R. (2007): *Cognitive Integration: Mind and Cognition Unbounded*. Dordrecht: Springer.

MORENO, A. & MOSSIO, M. (2015): *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Dordrecht: Springer.

MOSSIO, M. & BICH, L. (2017): What Makes Biological Organisation Teleological? Synthese 194(4), 1089-1114.

NEWTON, N. (2000): Conscious Emotion in a Dynamic System: How I can Know How I Feel. *The Caldron of Consciousness: Motivation, Affect, and Self-Organization*. John Benjamins Publishing Company, 91-108.

NICOLIS, G. & PRIGOGINE, I. (1977): *Self-Organization in Nonequilibrium Systems* (Vol. 191977). New York: Wiley.

PATTEE, H. H. (1988): Instabilities and Information in Biological Self-Organization. In: Yates, F. E. (ed.): *Self-Organizing Systems: The Emergence of Order*. New York: Plenum, 325-338.

PEZZULO, G. (2008): Coordinating with the Future: The Anticipatory Nature of Representation. *Minds and Machines* 18(2), 179-225.

RUIZ-MIRAZO, K. & MORENO, A. (2000): Searching for the Roots of Autonomy: The Natural and Artificial Paradigms Revisited. *Communication and Cognition-Artificial Intelligence* 17 (3-4), 209-228.

RUIZ-MIRAZO, K. & MORENO, A. (2004): Basic autonomy as a fundamental step in the synthesis of life. *Artificial Life* 10(3), 235-259.

SCHULKIN, J. (2003): Allostasis: A Neural Behavioral Perspective. *Hormones and Behavior* 43(1), 21-27.

STERLING, P. (2012): Allostasis: A Model of Predictive Regulation. *Physiology & Behavior* 106(1), 5-15.

THELEN, E. & SMITH, L. B. (1994): *A Dynamic Systems Approach to the Development of Cognition and Action.* Cambridge: MIT Press/Bradford.

THELEN, E. & SMITH, L. B. (1998): Dynamic Systems Theories. In: Bronfenbrenner, U., Morris, P., Damon, W. & Lerner, R. M. (eds.): *Handbook of Child Psychology*. 6th edition. New York: Wiley, 258-307.

VARELA, F. J. (1979): *Principles of Biological Autonomy*. Dordrecht: Elsevier.

VARELA, F. (1992): Autopoiesis and a Biology of Intentionality. In: McMullin, B. & Murphy, N. (eds.): *Proceedings of a Workshop on Autopoiesis and Perception*. School of Electronic Engineering, 4-14.

Organon F 25 (2) 2018: 244-264

Mathematical Models as Abstractions

LUKÁŠ ZÁMEČNÍK[1]

ABSTRACT: The paper concerns a contemporary problem emerging in philosophy of science about the explanatory status of mathematical models as abstractions. The starting point lies in the analysis of Morrison's discrimination of models as idealizations and models as abstractions. There abstraction has a special status because its non-realistic nature (e.g. an infinite number of particles, an infinite structure of fractal etc.) is the very reason for its explanatory success and usefulness. The paper presents two new examples of mathematical models as abstractions – the fractal invariant of phase space transformations in the dynamic systems theory and infinite sets in the formal grammar and automata theory. The author is convinced about the indispensability of mathematical models as abstraction, but somehow disagrees with the interpretation of its explanatory power.

KEYWORDS: abstraction – dynamic systems theory – explanation – formal grammar – idealization – mathematical model – Morrison – philosophy of science.

1. Introduction

I believe that in the current debate on the nature of scientific models the traditional question (typical for the semantic conception of scientific theories) of the relationship between abstract models and theories, has been

[1] Received: 8 March 2018 / Accepted: 23 April 2018

✉ Lukáš Zámečník

Department of General Linguistics, Faculty of Arts
Palacký University, Křížkovského 10
771 80 Olomouc, Czech Republic

e-mail: lukas.zamecnik@upol.cz

somewhat neglected. The current mainstream debate on the nature of scientific theories is commonly referred to as a pragmatic view of theories. This debate was launched primarily by Nancy Cartwright (1983; 1999) and Ronald Giere (1999; 2006), and can be summarized as an approach resigning to the description of scientific theory as an abstract structure with clearly defined relations between the individual components within this structure. Theory is simply conceived as a cluster of models that are appropriate to represent certain elements of the phenomena under investigation. Currently even the very idea of scientific theory is neglected in favor of the idea of scientific modelling (see e.g. Gelfert 2016, Zach 2017).

Scientific models, e.g. causal, non-causal (and plethora of their types), in this context are fruitfully investigated in terms of building the typology of models and in terms of important contributions to topics of scientific explanation and prediction (see e.g. Weisberg 2013). Yet I think this omits an important question central in the traditional philosophy of science. This question cannot be ignored, and is eventually testified by some texts of the proponents of the pragmatic conception of theories themselves, especially by Ronald Giere. In "Scientific Perspectivism" he modified his pragmatic conception of theories when, in addition to the introduction of data models, he conceded within the abstract models a definite place for principles (see Giere 2006, 61-62). However, Giere neglected the question of the nature of the nexus between principles and models.

The pragmatic view of theories works with models primarily as idealizations that are appropriate to represent a particular situation (for the researcher/scientist, see Giere 2006, 60, 62-63), given that they are similar to the data models investigated as "operationalized events /entities". The question of defining similarity (see Giere 2006, 63-67) as a sufficiently precise[2] concept will be shelved and we will focus on the view of mathematical models as abstractions.[3]

The aim of the study is to develop the concept of mathematical model as abstraction offered by Margaret Morrison. Her approach is inspirational because it overcomes the constraints imposed by the current concept of simplifying assumptions. This allows us to avoid the pitfalls of fictionalism

---

[2] Peter Smith accuses Ronald Giere of vagueness, see Smith (1998a, 253-277).

[3] We are aware about the debate concerning simplifying assumptions of scientific models, which are defined as abstraction and idealization, see Godfrey-Smith (2009).

and formalism (together we can call them mathematical utilitarianism), but also realism (or mathematical Platonism) in approaching mathematical models in the natural sciences (especially in physics).

However, although Morrison points at the peculiar position of mathematical models as abstractions, she does so only with the help of a relatively limited set of examples (linked to the renormalization group). In addition, she faces the problem of combining of unrealistic properties of explanatory models with an explanatory theory. This second problem is more serious because it can lead to a leap (rejected by Morrison) to the explanatory power of mathematics itself in relation to a natural science.

The first problem will be removed by presenting other two examples in which abstraction plays a crucial role. The first example is from the dynamic systems theory, the other example comes from linguistics, particularly from the field of formal grammars. The definition of the concept of abstraction by Morrison and the introduction of two new examples will be elaborated in the second and third sections of the study.

The fourth section will focus on the second issue of Morrison's approach. We will outline how to prevent the mentioned danger, through a close alignment of mathematical models as abstractions with their theoretical principles (which is also present in two new examples). The rehabilitation of the concept of the theoretical principle leads to the fulfillment of the explanatory potency of a scientific model (in our case: of abstraction).

2. Mathematical models as idealizations and abstractions

Morrison inclines towards pragmatic and pluralistic view of theories based on scientific models; she says that models act as autonomous mediators between theory and applications, or between theory and the world (see Morrison 2015, 20). However, in contrast to pragmatic-oriented variants of classical model-based views of theories (MOT) she fundamentally modifies the meaning of specifically mathematical models in this mediation by distinguishing their role according to whether they are abstractions or idealizations.[4]

[4] We have to notice that the way of using the term abstraction and idealization is slightly different from usage in context of simplifying assumptions.

Morrison states:

> (…) abstraction is a process whereby we describe phenomena in ways that cannot possibly be realized in the physical world (…); the mathematics associated with the description is necessary for modelling the system in a specific way. Idealization on the other hand typically involves a process of approximation whereby the system can become less idealized by adding correction factors (…) idealization is used primarily to ease calculation. (Morrison 2015, 20)

The last sentence reminds us of the classic MOT, which is characteristic of Ronald Giere where models are actually viewed as useful tools used to represent aspects of the world:

> What is special about models is that they are designed so that elements of the model can be identified with features of the world. This is what makes it possible to use models to represent aspects of the world. (Giere 2004, 747)

Morrison adds:

> In their original state both abstraction and idealization make reference to phenomena that are not physically real; however, because the latter leaves room for corrections via approximations, it can bear a closer relation to a concrete physical entity. (Morrison 2015, 20-21)

For this reason, models like idealization are favoured by most MOT supporters. Morrison, however, shows us that this view of the model and of its role in scientific theories are both extremely simplified. Morrison focuses on those cases of applying mathematical abstractions in models where these abstractions are not accessible to approximation techniques (see Morrison 2015, 21). Because, according to Morrison, these abstractions are necessary to depict and understand the behavior of physical systems, of which she says: "(…) the inability of standard accounts to capture the way mathematical abstraction functions in explanations" (Morrison 2015, 21).

Morrison comprehensively investigates the role of such abstractions, both in terms of their ability to provide general features of physical systems

(see Morrison 2015, 25-26), and, for her more importantly, in terms of their ability to provide: "(…) detailed knowledge required to answer causal questions" (Morrison 2015, 26).

The chief example chosen by Morrison is the dynamics of phase transitions:

> The occurrence of phase transitions requires a mathematical technique known as taking the "thermodynamic limit" $N \to \infty$ (…), we need to assume that a system contains an infinite number of particles in order to explain, understand, and make predictions about behaviour of real, finite system. (Morrison 2015, 27)

Morrison points out that this is not a kind of simplistic calculation but:

> (…) the assumption that system is infinite is necessary for the symmetry breaking associated with phase transitions to occur. (…) we have a description of a physically unrealisable situation (an infinite system) that is required to explain a physically realisable phenomenon (the occurrence of phase transitions in finite systems). (Morrison 2015, 28)

I believe that Morrison's fundamental insight into the exclusivity and indispensability of mathematical abstractions as a means of theoretical representation (see Morrison 2015, 29) is marred through excessive affinity of most of the cited examples to "emergent phenomena" (see Batterman et al. 2013). These are also closely related to phase transitions in connection with dynamic systems theory (hereafter DST, see section below). Moreover, when Morrison talks about the use of mathematical abstractions in biology, they occur in areas that are linked to DST (population dynamics). Taking into account other major Morrison texts, this becomes even more clear, because all the examples mentioned fall within the scope of scientific unification through universality (see Morrison 2013, 381-415).[5]

[5] Morrison even distinguishes three variants of unification of theories: through reduction, synthesis and on the base of universality.

The specificity of mathematical abstractions that even provide information[6] on the physical (or biological) system under investigation (see Morrison 2015, 55) is demonstrated through the example of a renormalization group (RG), which is used for the mathematical modelling of the dynamic system at critical points in phase transitions (see Morrison 2015, 57-67). These descriptions lead Morrison to DST and to the concept of universality:

> Diverse systems (…) with the same critical exponents exhibit the same critical behaviour as they approach critical point. In the sense they can be shown via RG to share the same dynamic behaviour and hence belong to the same universality class. (Morrison 2015, 70-71)[7]

Morrison talks about the ontological independence of the macro level of description at the micro level of description (see Morrison 2015, 74) and conveys the need to formulate a new concept of scientific explanation:

> Instead of deriving exact single solutions for a particular model, the emphasis is on the geometrical and topological structure of ensembles of solutions. Further explication of these aspects of RG methods allows us to appreciate the generic structural approach to explanation that RG provides. (Morrison 2015, 76)[8]

As evidenced by the citations, the whole discussion about the abstractions at Morrison concentrates on DST. In the first part of the next section,

---

[6] This is a rather vague part of Morrison's argumentation, where on the one hand it cannot be said that mathematics can provide an explanation of physical facts, but on the other hand it cannot be claimed that information about the physical system is included entirely in the physical hypothesis (and in specific conditions). Thus, mathematics acquires a specific status not only as a means for explaining but also as a co-constituent of information on the system under examination (see Morrison 2015, 55).

[7] Morrison also recalls the importance of power laws to describe regulatory parameters. She recalls a number of variants of these laws across disciplines (see Morrison 2015, 70). We should recall that they are also important in the context of quantitative linguistics (see Köhler et al. 2005).

[8] Significant similarity to Kellert's concept of qualitative prediction and description of geometric mechanisms (see Kellert 1993, 97-105).

abstraction defense will be used directly in the DST context on the level of application of fractal geometry. This shows the issues of phase transitions, critical points and the use of RG in another perspective, following the discussion by Stephen Kellert and Peter Smith.

To demonstrate that the importance of mathematical abstractions for understanding (or even for explanation) within scientific theories is not only tied to DST, we also provide a second part of the third section exploring the abstractions beyond DST and physics. We will focus on the importance of the mathematical model of an infinite set for automata theory and formal grammar.

3. In support of abstractions

3.1. Fractal geometry in dynamic systems theory

Dynamic Systems Theory (DST) is one of the central scientific concepts on which a large part of today's scientific applications, and new theoretical approaches rests. The debates of philosophers of science on the DST culminated in the 1990's and was predominantly formulated by Stephen Kellert (1993) and Peter Smith (1998). This theory, especially under the popularised name chaos theory, was in the focus of the philosophers of science for reasons connected with a pronounced relativisation of methodological criteria in the natural sciences. Foremost was the discussion about the revision of some important philosophical-scientific concepts – especially scientific law[9] in the context of the views of scientific theories and predictions within scientific explanations.[10]

The degree of change effected by DST, judging representatively on the basis of Kellert's and Smith's texts, is not too extensive and is well documented. Unfortunately, Smith's interesting idea of the importance of fractal geometry for the explanations of dynamic behaviour of the system, which is in a chaotic mode, has been largely unnoticed. We cannot reasonably

[9] Here we draw attention to Kellert's inspiration by Giere's studies of the 1980s.

[10] Today, the main debate is concerned with the issue of phase transitions and the associated universality of the description of phase transitions across various scientific ontologies. This also often involves the concept of emergence (see e.g. Batterman ed. 2013).

expound fractal geometry and its application to DST (see Peitgen et al. 2004). However, two aspects of this mathematical entity are essential for our purpose; the first is the infinity of the fractal structure and differentiating fractals from prefractals.

In the DST the concept of infinity was crucial. Its importance is appropriately summarized in the redefinition of Laplace's demon postulation. In order to allow unlimited predictions of the evolution of the dynamic system over time, in some cases (for certain control parameters) we need to know accurately all initial conditions of the dynamic system.[11] In short, Laplace's proverbial demon must indeed possess an infinite memory and omniscience.

This interpretation of the predictive constraint in DST is reflected in Kellert's concept of the transcendental impossibility of certain types of predictions (see Kellert 1993, 32-42). We refer to it here because we think it contrasts with the correct use of the mathematical model as an abstraction in the case of Smith. In the case of Kellert, an abstraction of infinite precision is used because the theory can demonstrate that for an arbitrary little inaccuracy of knowledge of the initial conditions, we always find (in the case of chaotic dynamics) the situation in which the error rate reaches the magnitude of the measured quantity. In other words we lose the ability (quantitative) to predict development of the system (sensitive dependence on initial conditions).

I believe that the abstraction of Laplace's demon with the infinite memory is inadequate, because the need to know all the details of a dynamic system is dispensable. From the empirical point of view, it makes no sense to think that the degree of inaccuracy is infinitely small, but it will be reflected in the final instance. The use of the infinity model is therefore in this case only idealization.

Similarly, when we use fractal geometry in many cases, it is enough to build on the knowledge of the most suitable prefractal without needing to work with the infinitely fine structure of the fractal. Analogous to Morrison's examples, the mathematical object of the fractal is used as an ideal object for only a certain aspect of creating a hypothesis (in relation to representation of the data model), to a certain level of accuracy (the number of iterations performed). Analogously, for example, because we know that

[11] Prigogine discusses this in "Order out of Chaos" (1984).

the sea border of Norway is not infinitely long, we do not need to revert to the molecular or even atomic level to describe the structure of its coast.

It seems that prefractals are therefore a good example of mathematical models as idealizations, as Morrison discusses. In this case, the mathematical object is not present in the theory or application of the theory as a whole, but only its appropriate scheme. Smith, however, also focuses on mathematical DST models that clearly correspond to how Morrison characterizes mathematical abstraction. Smith expresses the core of the problem in a simple argument:

> To summarize: we initially noted that
>
> (F) The chaotic behaviour in models like Lorenz's depends on trajectories getting pulled ever closer to a strange attractor with a fractal geometry.
>
> It has now been argued that
>
> (G) The evolving physical processes that chaotic dynamic models like Lorenz's are characteristically intended to represent cannot themselves exhibit true infinite intricacy.
>
> (F) and (G) together imply the conclusion that, at least in the typical case, the very thing that makes a dynamic model a chaotic one (the unlimited intricacy in the behaviour of possible trajectories) cannot genuinely correspond to something in the time evolutions of the modelled physical processes – since they cannot exhibit sufficiently intricate patterns at the coarse-grained macroscopic level. (Smith 1998, 41)

Still, according to Smith, we find cases (see Smith 1998, 41-45) where the mathematical entities of the fractal are generally used with the infinite depth of this structure, despite the empirical inadequacy mentioned above. Smith notes:

> We can live with this, treating it just another case of the way idealizing theories depart from strict truth, if we can find some compensating virtue – roughly, some story about simplicity to trade off against the empirical mismatch. (Smith 1998, 45)

And this simplicity Smith discerns:

> (…) if we stare at the infinite detail of e.g. the Lorenz attractor, we naturally think of it as an astonishingly complex object and then wonder how such a mathematical monster can legitimately get put to empirical work (…). But switch perspectives again, and think of the attractor as what is left fixed in place by a dynamics which stretches and folds phase space trajectories, and we now can see how the needed simplicity might get into the picture. For we could have a dynamic model which specifies relatively simple stretching-and-folding operations, yet (…) even very elementary stretches and folds can have infinitely intricate fractal invariants. (Smith 1998, 46)

My previous depiction of Smith's "new form of idealization" (see Zámečník 2012a, 699-703) now appears to correspond to the concept of abstraction used by Morrison. Similar to her examples, which work with models containing the mathematical infinity entity, we also need an infinite structure of the fractal. It is unavoidable that an explanation of the dynamics of the system is actually present in the form of an infinite intricacy of fractal invariant. The explanatory force of the theory depends on the fact that we work with the mathematical model as abstraction.

3.2. Infinite sets in formal grammar

Mathematical models like abstraction are also found outside the sphere of natural sciences. In linguistics, for example, they manifest themselves in the Chomsky hierarchy of formal grammars, which describes the path to transformational grammar. Even in this case, like Morrison's, we encounter a mathematical infinity, this time in the context of set theory. Again, it is not possible to fully capture the whole theory of the Chomsky hierarchy (see e.g. Partee et al. 1993, 559-561), but only to select the central aspects that will show the role of mathematical models as abstractions.

The fundamentals of Chomsky's transformation grammar are based on automata theory (see Partee et al. 1993, 431-435), when strings generated by individual types of grammars can be identified with strings accepted by individual types of state automata – for example, finite state automata correspond to regular grammars, pushdown automata correspond to context-

free grammars and Turing machines correspond to recursive enumerable grammars.

The role of mathematical models as abstractions appears in formal grammars in the very foundations of automata theory, where a crucial role is played by the fact that a power set made up of an infinite set of natural numbers is uncountable. For automata theory, the central aspect of set theory is the fact that one-to-one pairing cannot be done between an uncountable infinite set of real numbers and a countable infinite set of natural numbers.[12] This is because it is impossible to arrange the elements of the set of real numbers in a series, according to the given rules. For example, if we take real numbers from zero to one, we cannot find an algorithm that would lead to an endless series in which all the real numbers from this interval would be successively present (see e.g. Papineau 2012, 30-39).

Given formal grammar as a model of any grammatical system, although this model can be approached as idealization in the sense that formal grammar must be distinguished from the grammar of natural language,[13] formal grammar appears to be a non-reducible abstraction with respect to the above-mentioned aspects of set theory.

Partee states that, given that the means we take into account in the formal grammars for the characterization of language are countable infinite classes, it follows that there is an uncountable infinite number of languages that do not have grammar (in the above sense).[14] Therefore, there are such sets of strings that they cannot be characterized by finite means (see Partee et al. 1993, 433-434). The distinction between individual types of infinities, mathematical models as abstractions, plays a central role in defining the area of formal grammatical descriptions.

---

[12] The relationship between these sets is expressed in such a way that each member from the set of real numbers can uniquely pair with a member of the power set of natural numbers. Possibly stronger claims about the nature of the infinity of natural and real numbers are expressed in the continuum hypothesis.

[13] For example, the basic assumption that formal grammar, which is a suitable candidate for the representation of natural language grammar, must be at least slightly context sensitive (see Partee et al. 1993, 501-503).

[14] The argument resides, *in nuce*, on the fact that the language with the dictionary A can be defined as any subset of A^* (see Partee et al. 1993, 433). Assuming that A^* is countable infinite, power set $\wp(A^*)$ is uncountable infinite.

Here we may object to whether it is appropriate to consider abstractions and idealizations in the field of formal grammars if Morrison's and our examples are tied to the natural sciences, whereas here we are basically moving into a formal discipline that fundamentally draws on the set theory and algebra. We believe that this example is relevant and important because the importance of formal grammars rests, among other things, on their modelling role with respect to the natural language grammars (e.g the disputes about context-freeness and context-sensitivity of natural languages, see Pullum & Gazdar 1982, Schieber 1985).

Partee holds that languages characterized by final means show in their strings a pattern that distinguishes them from other strings in A^* (see Partee et al. 1993, 434). Although natural language grammars are much more complex than formal (and therefore we may speak about idealization), it is still essential that we approach natural grammars as sets of rules that simply have to be characterized by finite means. Thus, our concept of the natural language grammar (see also Chomsky's transformational grammar) is bound to work with the abstraction of infinity in the distinction of its countable and uncountable variants.[15]

In automata theory in connection with the Chomsky hierarchy, Turing's machine is of central importance, which accepts the strings generated by unrestricted rewriting systems (type 0 grammar), defining recursively enumerable languages. In concretizing the above, it is true that an infinite number of Turing machines can be uniquely coupled with natural numbers, that is, the Turing machines are countable infinite. Of course, it follows, according to this argument above, that there are uncountable infinite numbers of Turing's unacceptable languages (see e.g. Partee 1993, 505-523).

Morrison does not remain bound by physical examples when she claims that biology needs mathematical models like abstraction (see Morrison 2015, 40). In addition we can say that every comprehensive theory of grammar (not only formal) necessarily requires mathematical models like abstractions.

---

[15] We are aware that there is a large group of set theory critics with regard to the concept of infinity (see e.g. Vopěnka 1979). This text is intended, inter alia, to provide an apology of the concept of infinity in mathematics.

4. Why we cannot renounce our mathematical abstractions

In the previous two sections, we adhered to Morrison's position advocating the importance of mathematical models as abstractions, not merely as idealizations. We illuminated from a different perspective the role of DST abstractions and we documented that the use of abstractions is not limited by DST and the concept of universality. If we concede that the role of mathematical models is more complex than the pragmatic philosophy of science suggests, then the crucial question arises as to how to elucidate the relationship between mathematics and science.

Morrison puts this question in the above referenced book: "The interesting philosophical question is how we should understand the relation between this abstract structure and the concrete physical systems that this structure purportedly represents" (Morrison 2015, 22-23). This question is about the nature of the relationship between mathematics and physics. The question that Morrison poses elsewhere (see Morrison 2015, 55) is whether it is possible to separate mathematics and physics contained in physical theory.

The discussion in philosophy of science cannot be satisfied with merely spraying individual examples which can support a certain concept of the model. On the other hand, the two newly introduced examples of models designed as abstractions discussed above allowed the Morrison's concept to get rid of its excessive exclusivity in relation to a large but limited set of examples (the renormalization group). At the same time, we have facilitated the redirection of the main emphasis in conceiving abstractions from their role of means of representing phenomena to their role of explanatory theories. We believe that in both examples the binding of mathematical models as abstractions with theoretical principles is obvious (for more see below).

The position to be defended can be illustrated by the argumentation sketch as follows:

1. The inherent role of scientific models is to convey an explanation.
2. Explanation cannot be bound to purely mathematical entities, i.e. a mathematical fact cannot exclusively explain a natural fact.

3. Morrison does not present any concept of abstraction as a mathematical model that allows explanation, which does not contradict point two.
4. A common characteristic of the examples given in the third section is that they contain explanatory model (mathematical abstractions), because of relations of these models to theoretical principles.
5. The unrealistic nature of the model (with respect to point 4) does not prevent the model from participating in the explanation.
6. The concept according to which we define the preceding points is referred to as mathematical conventionalism.

We believe that point one of our argumentation frame does not require a special commentary. It is hard to imagine a science built purely on the base of models as appropriate representations of the system under study, without any possibility of defining their explanatory role. This task is based on the possibility of delimitation of the theoretical principles which the models are based on.[16]

Also, the second point does not need an extensive commentary to be supported, because we probably find only a few authors who would argue with it. Morrison deals with an analysis of several counter-examples, defined by Baker (see Morrison 2015, 50-57), and refuses the Baker's position. We agree with her rejection because we can say in terms of conditional reductionism that all explanations in natural science should ultimately be physical, but when accepting the mathematical explanation of the physical, we might accept the reduction of physics to mathematics.

At point two of our argumentation, it is particularly interesting why Morrison paid such attention. Morrison clearly stands away from a number of concepts of models (primarily she criticizes the concept inspired by Nancy Cartwright), one of the most important being fictionalism (see Morrison 2015, 85-118). We believe that she fails in clear declaration that her approach to mathematical models as abstractions cannot be interpreted just

[16] I thank to Ladislav Kvasz who once said in a discussion that the concept of models as representations of different systems without the knowledge of any unifying theory recalls the conception of ancient Egyptian science in which no theories existed, but only groups of applicable models/representations.

as a denial of point two. This problem is mainly related to the fact that it is not clear from the Morrison's argument what the physical information content (physical information) carried by the mathematical structure is.

We believe that what Morrison introduces when interpreting abstractions in the context of renormalization theories, i.e. in DST (used in other places as evidence of a specific method of unification in physics, see Morrison 2013), recounts more a sum of formal properties of a mathematical system that can be used to represent a real system. We, thus, believe the concept of Morrison's mathematical models as abstractions is similar to formalism.

In other words, it reminds us of a situation where we would argue that for example differential calculus carries information about the physical system and thus explains a class of dynamic phenomena. This example is pertinent because we also know that the assumption of differential calculus is unrealistic (at least in the context of a discrete structure dictated by quantum physics and a standard model of particles and interactions). But the differential calculus is not almighty, of course, the core of the explanation is ensured, being limiting to classical dynamics, by the Newton's laws of motion.

We claim that Morrison, as pointed out in point three of the argumentation sketch, does not have the tools to actually make models like abstractions able to participate in the explanation without the mathematical structure itself being responsible for the explanation.

The central point of our conception (in point four) is the assertion that what makes models as abstractions explanatory is their association with theoretical principles. Although we do not consider this statement to be controversial (like the one in point one and two), we believe that too little consideration is being given to it in today's professional discussions. Morrison's attempt to use the concept of physical information borne by mathematical abstraction is inadequate because the theoretical principle is an abstract entity that is empirically adequate construction created by a cognitive agent with regard to the unification of phenomena and the comprehensibility of the world. The world is here in agreement with Davidson and Searle (see Searle 2012, 199-200), a regulatory idea that is a condition of the intelligibility of our beliefs.

Smith's definition of the role of fractal geometry in the dynamic systems theory is a piece of evidence of how a mathematical model as

abstraction should be conceived in its relation to a theoretical principle. The definition of the attractor of a dynamic system assumes that we have a theoretical principle – in our case it is an abstract entity expressing the strech-fold process of transformation of the phase space. For a special set of dynamic systems, strange attractors can be shown to be empirically adequate models of real systems whose dynamics is in a chaotic mode. And for these cases, it is inevitable to connect the fractal geometry with its infinite structure with an attractor. A mere prefractal would not be an adequate model because it would not express all the essential features of the theoretical principle.

As we have already expressed above in relation to mathematical models as abstractions used in formal linguistics, the basic theoretical principle governing all formal approaches modelling the natural language is the requirement that sets of rules expressing the natural language grammar must be expressed by finite means. This means that when modelling a natural language, we must have a model as an abstraction that distinguishes countable and uncountable infinities.

Beyond the above (in the third section) mentioned, it can be reminded that, as part of generative linguistics built by Chomsky on the basis of the formal grammar hierarchy, we encounter the mathematical model as an abstraction. This model is an embedding operation, which is connected with the basic principle of transformational grammar – with the principle of recursion. The recursive procedure allows you to generate unlimited long strings (sentences) by applying the final set of rules. There is also the need to implement discrete infinity of recursive prescriptions in the model as an abstraction. Also, here the model would not be involved in the explanation if it stated that the number of recursive operations was finite.

In connection with the fifth point of the argumentation sketch, there is the clarification of how the theoretical principle can serve to explain when it has unrealistic properties. We believe that this fifth point is problematic and unacceptable for advocates of most forms of scientific realism. However, since we have already entered constructive empiricism, it is not our intention to refute or otherwise justify the non-adequacy of scientific realism.[17] Because of our rationale that the explanatory force depends on

[17] As the only consistent form of realism, we admit Searle's external realism, which we interpret transcendentally (see Zámečník 2012b, 25-30).

the relationship between the model and the theoretical principle, we do not have to thematize the issue of realism at all.

The concept of mathematical conventionalism that we stand for in the sixth point of the argumentation is compatible with constructive empiricism (following van Fraassen 2002) and conditional reductionism (following Kim 2005), which we have previously made a part of our argumentation. Constructive empiricism conceives theories (theoretical principles and models) as empirically adequate constructions whose relationship with the world can never be based on isomorphy or, more generally, similarity. Also, as for van Fraassen, we believe that the world is above all a regulatory idea, and that empirical adequacy is defined by empiricism as a stance that prevents some theoretical constructs from being conceptualized as structures (or objects) of reality, hence protect us against metaphysics (see van Fraassen 2002, 36-38).

Conditional reductionism is not necessary to define our conception of mathematical models as abstractions. It states that all explanations should be principally reducible to physical explanation. It is based on the view of physicalism that we can find in Jaegwon Kim, and whose platform is on the concept of functional reduction (assuming physical realization of function) (see Kim 2005, 161-170).

If constructive empiricism refers to the origin and nature of theories (theoretical principles and models), conditional reductionism refers to the principle form of explanation using these theories. We build mathematical conventionalism as a view that expresses the structure and the characteristics of theoretical principles. Mathematical abstractions are the means by which a limited cognitive agent imprints the structure into theoretical principles. Mathematical abstractions (of course, we have taken infinity only, in countable and uncountable forms) are the constructional rules of theoretical principles and hence models. We believe that the origin of mathematical conventionalism can be traced back through van Fraassen (1989) to Cassirer (1923) (and probably to Poincaré).

In science the role of mathematics in modelling is therefore genuinely structural, and we concur with Morrison that this involves both the use of idealizations and abstractions. Pace Morrison, however, we do not believe that the finding of universality (see Morrison 2015, 80-81) implies that the mathematical structure is strictly understood in its explanatory/understanding role independent of chosen theories (working across ontologies).

Morrison's examples chosen from DST obscure the possibility that this mathematical model as an abstraction (e.g. here RG) will be replaced by another, at a given moment, for a given empirical evidence, more appropriate.[18]

Mathematical conventionalism is a position that can be wedged between fictionalism (and formalism) and the transcendent conception of mathematical abstractions in relation to the world. It does not determine a scientific model to the role of useful fiction (or formal description tools) on one hand and of transcendent mathematical entity on the other. Mathematical conventionalism (along with constructive empiricism and conditional reductionism) simultaneously defines the space for the axiology of science, which stands for three fundamental epistemic values: empirical adequacy, unification of theories and the comprehensibility of the world (point-of-view invariance).

5. Conclusion

Here we have striven to demonstrate several examples of DST and to exemplify formal linguistics to support the concept of mathematical models as abstractions as conceived by Margaret Morrison. We have seen that the use of abstractions is not limited to DST. The lack of mathematical models as idealizations, which the utilitarianists favour, does not imply that the central role of mathematical abstraction is a proof of the validity of mathematical Platonism. Abstractions are the necessary equipment of our creation of theories because of the transcendental limits of our reasoning.

Pragmatic orientation in the philosophy of science has seduced us to forget the indispensability of models as abstractions for the creation of scientific theories not only in the fundamental research of theoretical parts of physics, but also in profane and for foreseeable widely applied theories. In conclusion, despite the mainstream, we can say that without mathematical models as abstractions, science would be merely a cataloguing activity. In

[18] See, for example, the versions of physical theories that the need of renormalization understand as the absence of a fundamental theory – a theory that simplifies the expression of unification (see Batterman 2013, 141-188, 224-254).

nuce: scientific hypotheses have an explanatory power in many cases precisely to the extent that the mathematical model is present as abstraction.

Acknowledgements

I express thanks to Colin Garrett for his help in revising the English version of the text, Dan Faltýnek for the idea of the role played by mathematical abstraction in transformational grammar and Martin Zach for introducing me into contemporary debate concerning scientific modelling in philosophy of science. This paper is a part of the project *Quantitative Linguistic Analysis in Selected Areas of Applied Linguistic Research*, No. IGA_FF_2018_011.

References

BAIN, J. (2013): Effective Field Theories. In: Batterman, R. (ed.): *The Oxford Handbook of Philosophy of Physics*. Oxford: Oxford University Press, 224-254.

BANGU, S. (2013): Symmetry. In: Batterman, R. (ed.): *The Oxford Handbook of Philosophy of Physics.* Oxford: Oxford University Press, 287-317.

BATTERMAN, R. (ed.) (2013): *The Oxford Handbook of Philosophy of Physics*. Oxford: Oxford University Press.

CARTWRIGHT, N. (1983): *How the Laws of Physics Lie*. Oxford: Clarendon Press.

CARTWRIGHT, N. (1999): *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

CASSIRER, E. (1923): *Substance and Function*. Chicago: The Open Court Publishing Company.

CHOMSKY, N. (1957): *Syntactic Structures*. The Hague/Paris: Mouton.

VAN FRAASSEN, B. C. (1989): *Laws and Symmetry*. Oxford: Oxford University Press.

VAN FRAASSEN, B. C. (1989): *The Empirical Stance*. London: Yale University Press.

FRENCH, S. (2014): *The Structure of the World*. Oxford: Oxford University Press.

GELFERT, A. (2016): *How to Do Science with Models: A Philosophical Primer*. Berlin: Springer.

GIERE, R. N. (1988): *Explaining Science: A Cognitive Approach*. Chicago: The University of Chicago Press.

GIERE, R. N. (1999): *Science without Laws*. Chicago: The University of Chicago Press.

GIERE, R. N. (2004): How Models Are Used to Represent Reality. *Philosophy of Science* 71, No. 5, 742-752.

GIERE, R. N. (2006): *Scientific Perspectivism*. Chicago: The University of Chicago Press.

GODFREY-SMITH, P. (2009): Abstractions, Idealizations, and Evolutionary Biology. In: Barberousse, A., Morange, M. & Predeu, T. (eds.): *Mapping the Future of Biology: Evolving Concepts and Theories.* Boston: Springer, 47-56.

KADANOFF, L. P. (2013): Theories of Matter: Infinities and Renormalization. In: Batterman, R. (ed.): *The Oxford Handbook of Philosophy of Physics*. Oxford: Oxford University Press, 141-188.

KELLERT, S. (1993): *In the Wake of Chaos*. Chicago: The University of Chicago Press.

KIM, J. (2005): *Physicalism or Something near Enough*. Princeton: Princeton University Press.

KÖHLER, R. (et al.) (2005): *Quantitative Linguistics. An International Handbook*. Berlin: De Gruyter.

MANDELBROT, B. (1977): *Fractals: Form, Chance and Dimension*. San Francisco: Freeman.

MORRISON, M. (2013): The Unification in Physics. In: Batterman, R. (ed.): *The Oxford Handbook of Philosophy of Physics*. Oxford: Oxford University Press, 381-415.

MORRISON, M. (2015): *Reconstructing Reality: Models, Mathematics, and Simulations.* Oxford: Oxford University Press.

PAPINEAU, D. (2012): *Philosophical Devices.* Oxford: Oxford University Press.

PARTEE, B. H. (et al.) (1993): *Mathematical Methods in Linguistics*. Dordrecht: Kluwer Academic Publishers.

PEITGEN, H.-O. (et al.) (2004): *Chaos and Fractals – New Frontiers of Science*. New York: Springer-Verlag.

PRIGOGINE, I. & STENGERS, I. (1984): *Order out of Chaos: Man's New Dialogue with Nature*. New York: Bantam Books.

PULLUM, G. K. & GAZDAR, G. (1982): Natural Languages and Context-Free Languages. *Linguistics and Philosophy* 4, No. 4, 471-504.

SCHIEBER, S. (1985): Evidence against the Context-Freeness of Natural Languages. *Linguistics and Philosophy* 8, No. 3, 333-343.

SEARLE, J. (2012): Reply to Commentators. *Organon F* 19, supplementary issue No. 2, 199-200.

SMITH, P. (1998a): Approximate Truth and Dynamical Theories. *The British Journal for the Philosophy of Science* 49, No. 2, 253-277.

SMITH, P. (1998b): *Explaining Chaos*. Cambridge: Cambridge University Press.

VOPĚNKA, P. (1979): *Mathematics in the Alternative Set Theory*. Leipzig: Teuber Texte.

WEISBERG, M. (2013): *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

ZACH, M. (2017): Axel Gelfert: How to Do Science with Models: A Philosophical Primer (book review). *Organon F* 24, No. 4, 546-552.

ZÁMEČNÍK, L. (2012a): Filosofická reflexe teorie chaosu. *Filosofický časopis* 60, No. 5, 685-704.

ZÁMEČNÍK, L. (2012b): External Realism as a Non-Epistemic Thesis. *Organon F* 19, supplementary issue No. 2, 25-30.

Organon F 25 (2) 2018: 265-278

Russell and the Materialist Principle of Logically Possible Worlds

JAN DEJNOŽKA[1]

In his review in this journal, Martin Vacek knows that the second edition of my *Bertrand Russell on Modality and Logical Relevance* has a difficult mission of revealing hitherto unsuspected major new dimensions in a great thinker whose work has already been investigated for over a century. Vacek has a very fine understanding of the book, and expresses only a few doubts. I am writing to explain away those doubts.

My first topic is a matter of general interest: the core thesis behind Vacek's main doubt. Vacek says, "The core of these [combinatorial] theories [of logically possible worlds] is a construction [out] of some distribution of matter throughout a spacetime region" (Vacek 2017, 264). Vacek unsurprisingly cites Armstrong and Quine, among others, in connection with this thesis. Let us call it the Materialist Principle of Logically Possible Worlds.

The principle is perfectly fine for materialists who hold that matter (or bodies, or physical events) is the only logically possible category, or perhaps even the only intelligible category. But an idealist who holds that minds (or ideas) are the only logically possible category, or perhaps even the only intelligible category, could only hold that possible worlds are

[1] ✉ Jan Dejnožka

Union College
2877 Burlington Street
Ann Arbor MI 48105, U.S.A.
e-mail: dejnozka@juno.com

different distributions of *minds or ideas* in spacetime. The idealist Leibniz, the father of possible worlds logic, does exactly that. And the dualist Descartes, who admits the logical possibility of disembodied minds, surely would hold that there are infinitely many possible worlds distributing *only matter* (I omit God), infinitely many distributing *only minds*, and infinitely many distributing *both*. And neutral monist David Hume finds bodies and minds equally unintelligible. He literally has no idea of them, since he has no impression of them. Surely Hume could only hold that talk of possible worlds is talk of different distributions of *impressions and ideas*. Thus idealists would have an Idealist Principle of Possible Worlds, dualists a Dualist Principle, neutral monists a Neutralist Principle, and so on. Thus the Materialist Principle begs the question against every metaphysic other than materialism. You would have to be a materialist to find it even plausible.

Vacek does not openly state that he is a materialist, or openly state that other categories are not even *logically* possible. But if he is not a materialist in this radical sense, why is he not finding the Materialist Principle obviously false? Why is he not finding even one single logically possible world in which there is no matter, but in which there is something else?

By "other category," I mean category of things that logically can exist even if matter does not. Note that "There is no matter, therefore there are no minds" and "There is no matter, therefore there are no Humean sense-impressions" are logical non sequiturs. Consider also categories such that worlds identical in distribution of matter, including any worlds identical in having *no* matter, logically need not be identical in distribution of those other categories. That is, even if we pretend it is logically necessary that if *something* mental exists, then *something* material exists, a possible world still could not be identified as a certain mere distribution of matter alone. Even an epiphenomenalist who believes that minds logically depend on bodies can admit different minds in possible worlds with identical bodies if the psycho-physical *laws* of causation of epiphenomena are different. And the mind-body supervenience thesis, that minds are identical if bodies are identical, logically can be true within each possible world *consistently with that*. But the supervenience thesis is logically contingent. It is both formally and intuitively a logical non sequitur. Thus

it would not actually be true in every possible world. In fact, it is obviously false for worlds with identical bodies (or no bodies) but different disembodied minds.

Why does Vacek think the Materialist Principle has anything to do with my book? I am not sure. I hope Vacek is not criticizing my metaphysic, because I never state my metaphysic, and I am not a materialist. I hope he is not criticizing Russell, because Russell never was a materialist. Perhaps Vacek could criticize us for failing to hold the Materialist Principle because we fail to be materialists. But to do that, Vacek would first have to prove that materialism is the true philosophy. And I hope he is not criticizing my interpretation for failing to report that Russell held the Materialist Principle, or, in the alternative, for failing to criticize Russell for failing to hold it. For Russell never was a materialist, and would never have held that principle. Thus I am at a loss on what the criticism is even about.—My views? Russell's views? My interpretation of Russell's views?

In the book, I discuss combinatorial theory of possible worlds talk only for the 1914–1918 Russell's logical constructionist / fictionalist phase. Russell eliminates all bodies, and all minds except his own, as logical fictions—as logical combinations of sensed and unsensed sensibilia. Sensibilia are mind-independent (note the unsensed sensibilia) and physically real, meaning they construct the physical world. But they are not tiny packets of matter or small bodies. They are purely phenomenal. Russell uses them precisely to *eliminate* matter and bodies. Thus for the 1914–1918 Russell, possible worlds talk is implicitly talk of combinations (distributions) of mere phenomenal sensibilia (and / or his own mind).

The 1919–1921 neutral monist Russell eliminates all matter and all minds, even his own, as mere logical constructions of noticed and unnoticed phenomenal events. This is the zenith of Russell's logical constructionism.

In both of these constructionist phases, Russell can and would admit constructed minds but no constructed bodies in infinitely many possible worlds consisting of noticed "wild particulars" (physically unlawful, i.e. physically uncorrelated, i.e. random noticed events).

Vacek's two doubts are fear of circularity and fear of incompleteness.

For most Russell scholars, the fear of circularity in Russellian logical constructionism would be the fear that the logical atoms and the logical compositions of things into atoms logically determine each other, and do so in the same way. If they determine each other in *different* ways, there is of course no circularity. But since these are all logical analyses, we may call this fear of mutual logical definability. Note that physical atoms are not percepts given in acquaintance, not even through an atomic microscope, but are themselves deep theoretical constructions, meaning remote from the periphery of observation. And for a consistent materialist, even percepts on the periphery of observation would have to be theoretical constructions out of physical atoms. See Quine (1975 and elsewhere).

Thus that might be a fear for a materialist. But Russell is no materialist. His logical atoms include sense-data (*sensed* sensibilia), and these are not constructions. Nor are they determined or defined by his logical constructions. They are given in acquaintance. Even unsensed sensibilia are not constructions. They are what is *predicted by* his logical constructions. They are the sense-data (sensed sensibilia) we *would* have if we *were* in such and such a location in spacetime, under such and such conditions. Thus there is no circularity of mutual definability in Russell.

There is a way out even for materialists. The problem of mutual definability is a foundationalist problem. But Quine admits a holistic web of scientific theory (Quine 1975 and elsewhere). This is not new. *Word and Object* starts with an epigram from Otto Neurath describing the rebuilt ship of Theseus (rebuilt using its own timbers), and cites Pierre Duhem.

And everyone has another way out: simply choose what to take as primitive. It has been known for over a century that conjunction and disjunction are interdefinable using negation. We can take either to define the other. Or we can take either the Sheffer stroke or the Quine dagger (Peirce arrow) as primitive and use it to define conjunction, disjunction, and negation.

That is what most Russell scholars would consider the fear of circularity in Russellian logical constructionism. But that is not Vacek's fear at all. Vacek says:

> Put even [more strongly], in order to metaphysically explain the goings-on in the actual world (explanandum) by means of recombinations

> (explanans) one has to posit a necessitation relation between the two. Since the relation is modal in nature, we deal with a circular analysis ([which] can be a reason for Russell's scepticism about modality as a fundamental or irreducible concept). (Vacek 2017, 265)

This overlooks a principal feature of Russellian logical analysis. Namely, Russellian analysis is always eliminative. The analysandum is always eliminated as a logical fiction. During his 1914–1921 constructionist phases, Russell expressly defines logical truth, and thereby implicitly analyzes logical necessity away, as merely being purely general truth that is true in virtue of its form. (We know he rejects modal entities, including modal relations.) For the analysis to succeed, the analysandum must be logically equivalent to the analysans. And this logical equivalence must be, in ordinary talk, logically necessary. But does that introduce circularity into Russell's implicit *analysis* of necessity? Does it mean that we must circularly "posit a necessitation relation" here? Not at all. The very analysis eliminates *all* talk of logical necessity*, including talk of its own logical equivalence relation* as logically necessary, and replaces it with talk of purely general truth that is true in virtue of its form. On this eliminative analysis, there is no necessitation relation at all, not even in the implicit analysis of necessity. The relation is just a logical equivalence that is purely general and true in virtue of its analytic form. Indeed, if Vacek's fear were correct, then every logical analysis of Russell's, even an analysis of a tree or stone, would imply a necessitation relation. But they all merely state logical equivalences for him. For his implicit analysis of logical necessity eliminates every logical necessity as a logical fiction.

Vacek's second and main doubt is fear of incompleteness. For most Russell scholars, the fear of incompleteness in Russellian constructionism would be the fear of incomplete analysis. Russell came to see the problem, and it led him to abandon constructionism. For while his constructions can be described in general terms, as temporal series of classes of sensed and unsensed sensibilia, they can never be specific logical analyses that can be true or false, since they can never even be completely stated. For each would have to describe infinitely many classes of infinitely many sensibilia, so as to analyze all the infinitely many ways the

ordinary thing in question logically could be ordinarily perceived. See my (2003, 177).

But even if Russell could have provided finite specific analyses, or alternatively, if we were satisfied that Russell's general logical analysis of the world is correct, and that it is a mere finite human limitation that we can never completely state a specific logical analysis since it would be infinitely long, a second and very different problem of incompleteness would remain. Quine calls it underdetermination.

Every scientific philosopher faces the problem of underdetermination regardless of her metaphysics or her logical analysis of the world, not just Russell. The later Russell, who anticipates Quine in epistemic holism, though not in holist theory of truth, is aware of it. In fact, the later Russell describes *two* problems: every empirical theory is logically consistent with infinitely many arbitrarily different interpretations of experience, such as that Venus is real only on Mondays, Wednesdays, and Fridays; and infinitely many theories predicting different future observations are equally compatible with any given finite set of past observations (my 1995, 175). Thus underdetermination is not a bad *consequence* of scientific theory to be avoided, but an ordinary, *pre*-philosophical fact that we must admit as given, and provide an account of, in our theory. If our account is adequate, then all is well. The later Russell and Quine use their respective sorts of holism to do this. If their accounts are inadequate, that is criticism, not scholarship. Thus this fear is criticism of Russell, not criticism of my Russell scholarship. It takes us away from my logic book and into philosophy of science (see my 1995; 2006). Here I think Russell is better than Quine. It is our evidence taken as a whole that makes it likely that Venus is real every day of the week, and that the future will be like the past, as opposed to the infinitely many arbitrary alternatives to those two statements, and regardless of whether truth is holistic. (I think Russell has good arguments against instrumentalist / coherence truth holism in the *Inquiry*, and I think they apply just as well to Quine.) But even an epistemic foundationalist can simply admit that underdetermination is an ordinary fact, and simply dismiss the arbitrary Venus and future alternatives as arbitrary.

That is what most Russell scholars would consider the fear of incompleteness in Russell. But that is not Vacek's fear at all. Vacek says:

> The worry from incompleteness arises as far as we recombine actual atoms only and omit possibilities of the[re] being merely possible atoms. Although I am not sure how strong the intuition 'there could be worlds with more matter' is, one can still back it up with a simple (transcendental) consideration: a world to which no individuals, worlds, or properties are alien would be an especially rich world. There is no reason to think we are privileged to inhabit such a world. Therefore any acceptable account of possibility must make provision for alien possibilities [cite omitted]. Dejnožka discusses alien individuals and alien properties in several places (pp. 52, 81, 166, 182) yet he, in my opinion, does not square MDL {1, 2, 3} with this (again, maybe disputable) possibility properly. (Vacek 2017, 265)

I have several comments.

First, why does Vacek assume that merely possible atoms must be literally nonexistent objects? Given that the existence of ordinary minds and bodies is logically contingent to begin with, why is it not enough to be able to *describe* mere possibilia in possible worlds *talk*? And why cannot Russell use his actual but unsensed sensibilia? We need to see not only arguments that there *are* nonexistent objects, but also arguments why they are *needed* to explain how there logically could have been more matter (or minds).

Second, can there be a "world to which no…*worlds*…are alien"? Can possible worlds contain other possible worlds? Certainly a possible world can include all the objects that are in some other possible world, but that is not the same thing.

Third, there can be no alien properties for Russell. His universals are in the realm of timeless being as opposed to possible worlds of existents. There can only be alien instantiations.

Fourth, by definition the actual world can contain no alien objects at all. An alien object is defined as one that is in at least one possible world but not in the actual world. That is, an alien object is defined as a merely possible object.

Fifth, no one possible world can include all alien objects, since infinitely many alien objects have contrary or even contradictory properties. The apple that could be in my hand cannot be both the purple one from

possible world 1 and the nonpurple one from possible world 2. Thus a world to which no object is alien is a logically impossible world. Thus we can "inhabit such a world" if and only if we can inhabit a logically impossible world. We can be and often are mentioned in *talk* of possible worlds, and in *talk* of impossible worlds, including ones that we both inhabit and do not inhabit. But for Russell, the only world is the actual world. For Russell, there are no merely possible worlds and no merely possible objects, but only talk of them.

Sixth, it is important to note that the unsensed sensibilia that are almost all of Russell's logical atoms are just as real or actual as the very few sensed ones. No sensibile is a merely possible object. Russell is very clear that they are all actual. It is just that we sense only a very few of them. Thus when we construct how we could have seen a certain apple under other circumstances, no merely possible sensibilia are involved. A different actual sensibile would be sensed. Sensibilia are not potential beings. They are actual beings that can be potentially sensed. Thus they are mind-independent. In fact, they are prior to and construct minds.

There are no alien sensibilia. Russell admits actual sensibilia of every possible sort everywhere at all times, in infinitely many different phenomenal "private perspectives" or "private worlds" that jointly construct the public world, to account for how we logically can perceive any ordinary thing anywhere in ordinary spacetime. This logically includes accounting for how we can perceive any logically possible *new* ordinary thing anywhere in ordinary spacetime. Sensibilia construct both existing and possible new ordinary things alike, with no need to admit alien sensibilia or nonexistent objects of any kind. We may call this "phenomenal plenum theory." Russell suggests this is like Leibnizian monadology, but without the monads, and with mind-independently real phenomena. See *External World*. This sixth comment also applies to unnoticed events in Russell's neutral monist phase.

Seventh, we must not be bewitched by the picture that a nonexistent object could somehow move out of a merely possible world and enter the actual one. For Russell that is not possible, not even as a mere change or reclassification of ontological status, since for him there is no such thing as a nonexistent object in the first place. And for Russell, following Leibniz,

no object can *move* from any possible world to any other possible world, not even in possible worlds *talk*. For possible worlds are *defined* by which objects are in them, and by their interrelationships. But one object can *be* in many worlds.

Eighth, however, new bodies logically can just pop into being. After all, bodies are logically contingent. Of course, they would have to be logically possible bodies. Thus, to use possible worlds talk, they would have to *have been* alien objects. They would still be in whatever merely possible worlds they were, but they would no longer be alien to ours. But that is only one way the world logically can increase in matter. Already existing bodies and their existing constituents logically can expand or contract in volume without increasing or decreasing in number. They logically can even pop out of being, thereby becoming alien objects. As Kant says, things can just fade away. Nor need an expansion involve adding alien "filler matter." Inverting Kant, the existing matter can just gain in intensity (mass) and / or just expand in volume. This shows that adding alien matter is unnecessary for the amount of matter to increase. It also shows that Vacek's argument is a non sequitur. For it is logically possible that when a new body comes into being, an existing body shrinks in volume, fades in intensity, or even passes away from being, transferring its matter to form the new body so that the total amount of matter remains the same, or even decreases. As Quine might say, there logically can be compensatory adjustments. (That could even be kept up indefinitely, if there is an infinite amount of matter.) But for Russell, there is no matter in the first place. He would construct material expansion as replacement of smaller actual sensibilia with larger actual sensibilia, material contraction as the reverse, and new bodies in terms of new sensings of actual sensibilia.

Vacek seems to think that if there is no alien matter, then the law of conservation, that matter (mass-energy) can be neither created nor destroyed, is a *logically* necessary truth. Most people think it is a logically contingent physical law at best. Vacek is implicitly wondering how Russell can admit alien bodies in different possible worlds, as if for Russell, the very same amount of matter must be in all logically possible worlds. (And even if the amount could not change, even the ancient atomists could admit

recombinations of atoms into different bodies.) Perhaps Vacek is mistaking actual matter for Aristotelian potential matter, or even for Parmenidean being. If so, there go two more questions begged.

If there is no alien matter, so that for Vacek the amount of matter cannot change, so that for Vacek the amount of matter is logically fixed, exactly how much matter does Vacek think it is logically necessary for there to be? The actual amount in the actual world? How convenient! But what amount is that, and why is it exactly that amount? Why not 1% occupation of spacetime? Why not 25%, or 50%? The only answer that makes any sense would be 100% or total occupation of spacetime (material plenum theory). But is not empty space logically possible? Can there not be even a single vacuum, even for a moment? Cannot some possible worlds have more empty space than others?

Could there not have been one more apple, or even just one more electron, than there is? Russell's answer would be a hearty yes, spacetime logically could have been filled with many more things. And it logically could have been filled with many less, or even with nothing. But for Russell, all that is mere *talk* of what is possible. There are no alien objects out there in other possible worlds, objects that could somehow become real. Kripke for one would heartily agree. Here we may ask once again if Vacek is criticizing Russell or my scholarship of Russell.

Russell accepts the law of conservation, but only as a logically contingent law of physics, that is, as a law that is *not* true (and that is also not false) in all possible worlds. Russell does not even accept it as *causally* necessary. Russell is Humean and rejects causation. For Russell, a scientific law is no more than a uniformity of nature. Thus for him the law of conservation would merely assert a uniformity of the amount of matter across time in the actual world. Whether the universe is expanding or shrinking, or will eventually shrink, and the roles of entropy and of conservation of mass-energy in this, are scientific issues beyond the scope of this paper. I think we simply do not know. But surely all the rival scientific theories are at least logically possible. And Russell always kept up on the latest science. If he had lived longer, he might well have come to question or reject the law of conservation on purely scientific grounds.

Vacek does not openly state that the existence of matter must be logically necessary for Russell, much less that the exact same amount of matter must be in all possible worlds for Russell (the latter thesis implies the former), if Russell does not admit genuine "merely possible atoms" (Vacek 2017, 265). But it sounds for all the world like he thinks that this non sequitur is valid. And if he does not, then why he would criticize Russell on alien objects like this? And if Vacek does hold those views, then he must think that Russell cannot even admit a possible world with no matter. Of course, the view that one possible world has no matter and all the rest logically must have exactly the same amount of matter is, if anything, even more absurd, since the empty world would show that matter is logically contingent.

The 1914–1918 Russell has no problem of alien matter. He admits no matter at all. His sensibilia are logically contingent (MDL {1}, the level of logical atoms). Thus his constructions of minds and bodies are logically contingent (MDL {2}, the level of constructions). Thus he implicitly holds a logical constructionist version of the Dualist Principle for descriptive talk of possible worlds including some with only constructed bodies, some with only constructed minds, and some with both, and can talk of an empty world as well (MDL {3}, the level of language). Matter is eliminated as a logical fiction. All logical atoms are already there in the actual world, whether they are sensed or not. The actual world logically could not be richer in phenomenal logical atoms (sensibilia). It is a phenomenal plenum. Existing bodies, new bodies that come into being, and bodies that pass away from being are all logical constructions based on which logical atoms we sense. And while two material apples cannot both exist in the same spacetime region, infinitely many sensibilia can and do exist in the same constructed spacetime region as different sensible aspects of infinitely many constructible apples.

The 1919–1921 Russell implicitly holds the Neutral Monist Principle. His phenomenal but real events, some noticed (i.e. members of some set-constructed mind) and some not, are logically contingent (MDL {1}). Thus his constructions of minds and bodies are logically contingent (MDL {2}). Thus he can in principle describe infinitely many worlds that construct (1) only bodies, (2) only minds, (3) both, or (4) neither (if the events are too

few and / or too wild); and he can describe a world with no events as well (MDL {3}). Thus he also implicitly holds a constructionist version of the Dualist Principle. All worlds of type (2) and infinitely many of type (3) will have disembodied minds. But Russell believes our evidence is that the actual world has only embodied minds, i.e. only constructed minds that constructionally overlap with constructed bodies in the right way.

Vacek knows I explain several ways in which Russell can admit talk of alien objects. Vacek doubts these ways. I am perfectly satisfied with all the ways I list, and invite the reader to look up "alien objects" in the index. Of course, all the ways use *descriptions*. Russell is already using descriptions of nonexistents in his famous 1905 "On Denoting."

Thus it is very easy to "square MDL {1, 2, 3} with" alien possibilities (Vacek 2017, 265). Talk of alien objects is always done by descriptions. Descriptions always belong to MDL {3}, the level of language. And all statements via descriptions that alien objects exist are *false* general statements. This is clear as early as "On Denoting." There Russell analyzes "The present King of France is bald" as a false complex general statement. Talk of the present King of France is *talk of an alien object*. For the present King of France logically could exist, but does not. All such statements are false for Russell because for him there are no merely possible objects. That is because of his famous "robust sense of reality." That robust sense of reality is why Russell says "in some places" that possible worlds talk is "mere 'phraseology'" (Vacek 2017, 262). Phraseology, of course, belongs to MDL {3}, the level of language.

If Vacek doubts that Russell can talk about the present King of France using his theory of descriptions, or that a main point of the theory is to refute Meinong's admission of nonexistent objects (including both alien objects and logically impossible objects), or that Russell can legitimately assert that the actual world logically could have had more or less matter than it does (constructed or not), that is criticism of Russell, and not of any Russell scholarship I know of. But if I may humorously paraphrase Russell's famous scope distinction example of the yacht in "On Denoting," where a guest said he had thought the yacht was larger than it was, and the owner replied that no, his yacht was not larger than it was, even Russell would agree that if we use the owner's scope, then the actual world could

not have more matter than it does. But scope distinctions concern propositional attitudes, not metaphysics.

Vacek has produced a basically perfect description of my book, and expresses only a few doubts. Unfortunately, the Materialist Principle has nothing to do with Russell, and does not even seem to be true. And Vacek's doubts seem to be about Russell, not about my book. Certainly they have nothing to do with the book's success in revealing that there *are* major modal and relevantist dimensions in Russell's philosophy, regardless of whether his views are *correct*. Quite the opposite. Insofar as Vacek is doubting that Russell's modal views are correct, he is agreeing with me that Russell does have modal views.

The Materialist Principle logically entails neither constructional circularity, nor constructional incompleteness, nor even the law of conservation of matter (neither as a law of logic nor as a law of physics). Those are all non sequiturs. And except for materialists who reject even the logical possibility of other categories, the Materialist Principle is obviously false in the first place. Russell never held it, and would reject it. And (so) it has nothing to do with his logical constructionism. In fact, Russell rejects materialism throughout his career. He admits at least probable physical objects both before and after his constructionist phases, but he never admits physical objects alone.

I thank Vacek for a very fair-minded, kind, and even generous review.

References

DEJNOŽKA, J. (1995): Quine: Whither Empirical Equivalence? *South African Journal of Philosophy* 14, No. 4, 175-182.

DEJNOŽKA, J. (2003): *The Ontology of the Analytic Tradition and Its Origins: Realism and Identity in Frege, Russell, Wittgenstein, and Quine*. Lanham, MD: Littlefield Adams.

DEJNOŽKA, J. (2006): Observational Ecumenicism, Holist Sectarianism: The Quine-Carnap Conflict on Metaphysical Realism. *Philo* 9, No. 2, 165-191.

DEJNOŽKA, J. (2016): *Bertrand Russell on Modality and Logical Relevance*. Second Edition. Ann Arbor, MI: CreateSpace.

QUINE, W. (1975): *Word and Object*. Cambridge, Mass.: The M.I.T. Press.

RUSSELL, B. (1905): On Denoting. *Mind* n.s. 14, No. 56, 479-493.

RUSSELL, B. (1929): *Our Knowledge of the External World*. London: George Allen & Unwin. Revised ed. 1929, 1st ed. 1914.

RUSSELL, B. (1940): *An Inquiry into Meaning and Truth*. London: George Allen & Unwin.

VACEK, M. (2017): Review of Dejnožka (2016). *Organon F* 24, No. 2, 261-266.

Organon F 25 (2) 2018: 279-287

Zuzana Rybaříková: *The Reconstruction of A.N. Prior's Ontology* Univerzity Palackého v Olomouci, 2016, 134 pp.[1]

This is Rybaříková's dissertation in book form, which she defended at the Palacký University of Olomouc under the supervision of Jan Štěpán. In the interest of full disclosure, Petr Dvořák, who is my own dissertation advisor, was on her dissertation committee. Rybaříková's primary goal is to trace the development of Prior's thought vis-à-vis some thinkers who influenced him and with whom he disagreed: "[M]y study is primarily a historical work focused on the evolution of Prior's ontological views" (p. 18). Contrary to what the title may suggest, no systematic reconstruction of Prior's ontology is attempted. Indeed, it would seem that in Rybaříková's view, no such reconstruction is possible—at least not a consistent one—for Prior "did not present one consistent concept of ontology" (p. 118). In the first part of this review I present some of the logico-ontological theses which Rybaříková ascribes to Prior. In the second part I comment on what I think are some of the strengths and weaknesses of Rybaříková's work.

Prior was an unabashed nominalist in the sense that he thought that to be or to exist is to be a concrete thing. He also thought that propositions, possible worlds, properties, and moments of time are not concrete things, and therefore, given his nominalism, they don't exist (pp. 12*ff*). Though moments of time do not exist, the present is real, but the past and the future are not. One might think that the present is real only if at least one moment of time exists—namely the present one—and therefore Prior's nominalism conflicts with his presentism. But, for Prior, the present is real not in the sense that the present moment exists, but rather that the only things which exist are the ones which exist presently (p. 16).

To be or to exist is indeed to be the value of a variable if and only if the variable in question ranges over concrete things. If the variable in question ranges over propositions, properties, moments of time, and the like, then to be or exist is not to be the value of a variable. Thus, one may freely slide between

[1] ✉ Derek von Barandy

Department of Logic, Faculty of Arts
Charles University in Prague, Celetná 20
116 42 Prague 1, Czech Republic

e-mail: derek@logici.cz

> "it's possible that *p*" and "for some possible world *w*, *p* is true",
>
> "*p* was true" and "for some past time *t*, *p* is true at *t*",

as well as

> "*p* is a world proposition" and "*p* is true, and for any proposition *q*, if *q* is true, then necessarily, if *p* is true then *q* is true",

and the like without expressing theses which imply that a possible world exists, or that a time exists, or that a proposition exists (pp. 46*f*; 53*ff*; 87). Thus, for example, Prior may both agree with Quine that propositions don't exist, and yet disagree with Quine in maintaining that, in many cases, it is not sentences, either written or spoken, but propositions, construed as the "contents" (p. 85) of one's thoughts, which one believes, doubts, knows, and so on. And what are propositions? Following F. P. Ramsey, they're "logical constructs". Against Frege, however, they are human inventions and therefore are "dependent on the human mind" (p. 84).

If we may quantify over propositions, times, and worlds, and yet deny that they exist, may we also do the same for concrete things which don't presently exist but did exist (e.g. Napoleon) or will exist (e.g. someone's future child)? No and no.[2] From what I gather from Rybaříková's exposition (pp. 101*ff*) of Prior here, the central problem seems to be as follows: We may quantify over Napoleon only if he is the value of a variable. But since Napoleon is (or was) a concrete thing, he is the value of a variable only if he exists, and since everything which exists is that which exists presently, it follows that Napoleon exists presently. But Napoleon doesn't exist presently. Ergo, etc.

Any attempt at quantifying over future concrete things is plagued by a similar problem. However, this isn't a drawback for Prior because, unlike past individuals for whom or for which there are "state-able facts" (p. 111) (e.g. that Napoleon was an emperor), there are no future facts about any concrete thing, either that it exists

---

[2] In some passages Rybaříková contradicts what I'm about to say. For example, we're told that "a distinct feature of Prior's presentism was that he allowed quantification over objects which are not present" (p. 16) and that Prior "admitted that there were also facts about entities which had not begun to exist yet" (p. 103). However, since neither statement is accompanied by further comment, and because they seem to contradict the main tenor of Rybaříková's account, I'm not sure what to make of them. On p. 55 there's a hint as to how Rybaříková's Prior may avoid contradiction, for we're told that, for Prior, free variables "can stand for non-existent entities" and that both modal and temporal operators may "bind variables which stand for actually non-existent individuals". However, here again Rybaříková doesn't elaborate.

(or not), or that it is such and such (or not). This is just a consequence of Prior's brand of indeterminism with respect to the future.

Rybaříková's book is well structured. Each of the main chapters is on a certain logico-ontological theme in which Prior's views (or lack thereof, if Prior's didn't have a settled position) are presented after an extended exposition of some views of some thinkers who influenced him or with whom he disagreed. Here I shall list the themes and the names of the thinkers who are prominent in Rybaříková's discussions:

Chapter 2: Possible Worlds and Time Instants: Jan Łukasiewicz, Wittgenstein, and Carew Meredith
Chapter 3: Quantification: Quine and Stanisław Leśniewski
Chapter 4: Propositions: Quine, Frege, and F. P. Ramsey
Chapter 5: Names and Individuals: Leśniewski and Russell

Unfortunately, Rybaříková's characterization of a thinker's view isn't always accurate. In discussing Frege on one version of the problem of the substitution of identicals in intensional contexts as it appears in his "On Sense [*Sinn*] and Reference [*Bedeutung*]" (1948, §§34*ff*), Rybaříková says that Frege would say that the following sentences

[a] … evolution is based on natural selection.

[b] … evolution runs at the level of genes.

are not inter-substitutable in a context such as

Darwin believed that …

For, Rybaříková tells us, though [a] and [b] have "an identical sense" (p. 75), from

Darwin believed that [a]

It doesn't follow that

Darwin believed that [b]

But it's the other way around: Frege would say that [a] and [b] have the same referent (*Bedeutung*)—the True (or so we'll suppose)—but not the same sense (*Sinn*), as [a] and [b] express distinct thoughts.

In several places Rybaříková's wording invites the reader to confuse variables with their values. Here are some examples:

> "Prior differentiated among individuals bound by a quantifier, which refer only to some of the existent individuals, and among those which are free and can also stand for non-existent individuals." (p. 55)
>
> "Ramsey maintained that only individuals referred to existent entities." (p. 58)
>
> "[…] according to Prior there are also quantifiers [… [that] …] do not require the actual existence of entities which are bound by them." (p. 63)

In some cases Rybaříková's exposition of Prior is underdeveloped. Rybaříková reports (p. 101) that Prior claimed that the following schema

$\varphi y \supset \exists x\, \varphi x$

has some false instance. The counter instance cited is:

> Alexander rode Bucephalus ⊃ Some [presently existing] *x* is such that Alexander rode *x*

and we're told that this is a counter instance because its antecedent is true, but since Bucephalus no longer exists, and presumably no currently existing thing is such that Alexander rode it, its consequent is false. Now, by my lights, in order to have an instance of the schema's antecedent, we need an expression with a free variable, but any straightforward translation of "Alexander rode Bucephalus" would employ only constants—one for Bucephalus and one for Alexander. I suppose, though, that we're to assume that Bucephalus is the value of the free variable in, say,

> Alexander rode *y*

In which case the counter instance would be

> Alexander rode *y* ⊃ Some [presently existing] *x* is such that Alexander rode *x*

So far, so good. But, in Rybaříková's understanding of Prior's view, Bucephalus is the value of '*y*' only if he presently exists (given that Bucephalus is (was) a concrete thing), in which case, in accordance with the schema and on Prior's own terms,

> Some [presently existing] *x* is such that Alexander rode *x*

So I don't see how Prior could consistently maintain that

> Bucephalus is the (or a) present value of '*y*' in 'Alexander rode *y*'

and it's false that

Some [presently existing] x is such that Alexander rode x

and there's not much in Rybaříková's exposition which would help the reader see that, appearances to the contrary, Prior isn't inconsistent here, or that the reader should be confused because Prior, at least it in her view, is inconsistent.

Aside from these drawbacks, someone looking for a general overview of Prior's views on some fundamental logico-ontological issues, especially in relation to the thinkers mentioned above, as well as the nuts and bolts of some logics for which Prior was either an important innovator or sole inventor, will find it in Rybaříková's book.

Derek von Barandy

References

Frege, G. (1948): On Sense and Reference. Transl. by Max Black. *The Philosophical Review* 57, No. 3, 209-230.

Zuzana Rybaříková: *The Reconstruction of A.N. Prior's Ontology* Univerzity Palackého v Olomouci, 2016, 134 pp.[1]

The study of Zuzana Rybaříková is presented as predominantly an historical work. It is mainly focused on ontological ideas of Arthur Prior. She tries to discover some influences and to trace Prior's ideas in debates with those contemporary thinkers that had significant impacts on his development. Her particular interest seems to be in Prior's connections with members of Lvov-Warsaw school.

[1] ✉ Vladimír Marko

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

e-mail: vladimir.marko@uniba.sk

The study on Prior is divided into four sections: a character and origins of the concept of possible world; a way of handling non-nominal approach to quantification and interpretation of various prefixes as quantifiers; an interpretation of the concept of proposition; a comparison of the concept of names to the concept of individuals.

In the opening part of the work, Rybaříková tries to situate some basic assumptions at Prior's philosophical background. An appropriate definition of nominalism that could be ascribed to Prior, according to her, consists in complete denying the abstract entities. The idea is based on a distinction taken from *Objects of Thought* (Prior 1971, 3: "an object of thought is 1) what we think or 2) what we think *about*"). The position is further identified by Simons' fourfold demarcational definition and diagnosis of nominalism in Poland. Although this demarcation is usual, this idea was frequently criticized (see, for example Hugly & Seyward 1996, Ch. 2). Another basic point is Prior's nominalistic approach toward intensional logic and systems of modal logic – it consists of his denying the real existence of *possible worlds* and *possibilia*. The last point is his defence of presentism (Prior 1968, Chs. 1 & 12; 1970). It is here interpreted as a position linked with temporal realism – the conception that enables him quantifying over objects that are not present. This last formulation is only briefly exposed and seems to need some further elucidation for its stronger cogency.

Chapter 2 – devoted to possible worlds and time instants – ascribes sources of some Prior's ideas (following Suszko's interpretation) to Wittgenstein: possible worlds consist of propositions while world-proposition contains a conjunct of true propositions about the world. The development of formal systems of Prior is related to influence of Łukasiewicz and to his known attacks on determinism. Prior was well acquainted with works of Łukasiewicz. Soon after Łukasiewicz's death Prior took part in work of Meredith, Łukasiewicz's student who tried to formulate newly introduced values dealing with contingency and truth in a world alternative to the actual one. The criticism of Meredith's results on Łukasiewicz's work and some recognized outcomes of the three-valued logic later enabled Prior to independently develop his own systems of logic abandoning many valued logics. This step corresponds to his study on themes from history of logic and on discussions related with Diodorus' *Master Argument*. Here, for the first time, he explicitly expressed the connection between time and modality. Rybaříková's discussion of this issue consists of a too brief sketch of his ideas – there are many places Prior devoted to the defence of his conception of contingency and he frequently analysed the theme in his works with due care (for example, chapter 13 of his 1968), and he sometimes called these systems a (modal) logic of contingent existents.

According to Rybaříková, while Prior takes possible worlds useful as a tool, he never fully elaborated on the problem of the nature of non-existent individuals and the definition of possible state of affairs. For him, these questions remain open. Later, under the influence of Kripke and in accordance with his own indeterministic orientation, Prior introduced the concept of possible world in connection with the branching time structure with fixed past and open future. As it is known, the idea was based on recent researches that led to a structure expressed by Hamblin's lattices. Further development of temporal calculus in the book is characterized as corresponding to McTaggart's A- and B-series, respectively – where A-logical systems are related to his *presentistic* representation of time while U-calculus (*l*-calculus of *later than*) relates to B-series. Reduction of B-logical systems to A-logical systems of K_t led to some sort of hybridization of modal logic extensions, where a new sort of propositional symbols, called nominals, are linked to exactly one point (the idea should be ascribed, according to Rybaříková, p. 39, to an impact that came from Łesniewski's *Prototethic*).

Chapter 3 is devoted to Prior's theory of quantification. Here, modal, temporal and some other types of operators should be interpreted as quantifiers. The chapter consists of a longer introduction related to the confrontation of Quine, Ramsey and Łesniewski on nominal vs. non-nominal interpretation of quantifiers and of the questions regarding existential import, ontological commitment and *Barcan form*. Prior's response to the debate is characterized by attempting to make visible the distinction between existent and non-existent entities by introducing different kind of variables.

Chapter 4 is devoted to the ways Prior dealt with the notion of proposition. The influence of Ramsey and an inspiration taken over from studies on history of logic (especially Mates' accounts regarding Stoics logic and the logic of some medieval authors) inspired him to restate some of Quine's thoughts and take into consideration an ancient idea that the truth value of proposition is not fixed and can change throughout time (Prior 1968, Chs. 1, 13). Unlike Frege (interpreted here as adopting an indexical theory of proposition in which each sentence is unique regarding the circumstances of its utterance), Prior held, according to Rybaříková (p. 73), that the sentence is the same regardless of the circumstances in which it is uttered. Similarly as in the previous chapter, she tries to situate Prior's position in comparison with Frege's theory of propositional attitudes, Quine's rejection of intensionality and Ramsey's predicate analyses of proposition. Prior, preferring Ramsey's approach, held that a proposition is a logical construct and, at the same time, he criticized the view that propositions are genuine objects independent of the human mind. Rybaříková's final debate on his position that he left unelaborated and seems to be far from consistent is based exclusively on his posthumously published manuscripts *Object of Thought.* The

genesis of his opinions and some of his confrontations regarding the subject, however, could be found in many other places (for example in the opening parts of his posthumously published *The Doctrine of Propositions and Terms*).

Chapter 5 is devoted to Prior's notes on names and individuals mainly with respect to his studies on tense logic. The introductory part considers Russell's and Łesniewski's ideas on the subjects as a starting and explanatory point for forming Prior's own position exposed especially in his System Q (Prior 1957, Ch. vii) – consisting of the Russellian ΣT_1 (that admits logical proper names only for objects that have present existence), ΣT_2 (where any expression that makes a statement at any time makes a statement at all times) and ΣT_3 (that emphasizes difference between *the strong 'the'* and *the weak 'the'*, as proposed by Łesniewski (Prior 1957; Ch. viii)). The difference between these systems is exposed mainly with respect to Russell (and his differentiation between the definite and the indefinite article) and Łesniewski (with respect to the article-free use in Polish language, since he does not retain this difference, leading thus to different sorts of understanding of the verb "is"). The discussion continues with comments on Prior's rejection of some theorems of modal and predicate logic – with his interpretation of the *Barcan form* and with some peculiarities of the systems included in Q with respect to his temporal ontology where some problems arise in intensional interpretation of ΣT systems. For Prior, an advantage of Q could be obtained from ΣT_3 where some specifically defined individuals could be appropriately and successfully identified even in intensional context. The system was never fully axiomatized by its author, although he developed and improved some of its aspects in his latter works. The sub-chapter on *identifiable individuals* (and on Wilson's question "What would the world be like if Julius Caesar had all the properties of Mark Antony and Mark Antony had all the properties of Julius Caesar?") deals with Prior's comments related to the difference between truths about individuals that are *necessary* and those that *already happen* or are *possible with respect to some time while with respect to some other time impossible* (in the sense of unalterability of the actual state of affairs). The topic is further discussed in the following sub-chapter *Opposite numbers* in which epistemic reasons and the non-existence of two alternatives precludes us to comply with the future identity in the same way we deal with the actual one.

The book ends with a short concluding remarks. At this point, we would expect summarizing accounts related to the basic theme of book, namely the reconstruction of Prior's ontology. It is certainly hard to systematically grasp some work that is left unfinished by its author but some key points or concluding remarks related to the genesis of his opinions would be naturally expected.

There are redundant references at some places (an example is on p. 23: "in further section", "in further part"). Furthermore, the footnotes are hard to follow since they are printed in extremely small and paled font.

The assumption in the background of this work is that the reader is already acquainted with Prior's logical and philosophical contributions to some extent. However, although his texts are provocative to modern reader and are written in a quite stimulating manner, Prior is not so frequently discussed an author (this is usually explained in terms of his preference of the Polish symbolic notation). Given this, there was an opportunity to write this book for less informed readers; in such a case, however, many places in which technical aspects of Prior's systems are analysed should have been exposed in more details for the sake of transparency of his ideas and better readability of the text. Prior communicated with many persons of his age and was involved in many debates with those whose results have marked the development in many areas in logic and philosophy. Since the book is presented primarily as an historical study by reflecting mostly a dominance of Polish influences on Prior's work, it partly ignores some other important discussions in which Prior was involved and other influences that deeply or substantially affected him.

Beside the last remark Rybaříková's book is a rare and worthy attempt at throwing some light on thoughts of the philosopher who deserves our attention since in many realms he marked his own epoch and strongly influenced contemporary logic and philosophy.

Vladimír Marko

References

Prior, A. N. (1957): *Time and Modality*. Oxford: Oxford University Press.

Prior, A. N. (1968): *Papers on Time and Tense*. London: Oxford University Press.

Prior, A. N. (1970): The Notion of the Present. *Studium Generate* 23, 245-248.

Prior, A. N. (1971): *Objects of Thought*. Edited by P. T. Geach P. T. and A. J. P. Kenny. Oxford and Toronto: Oxford University Press.

Prior, A. N. (1976): *The Doctrine of Propositions and Terms*. Edited by P. T. Geach P. T. and A. J. P. Kenny. Amherst: University of Massachusetts Press.

Simons, P. M. (1998): Nominalism in Poland. In: Srzednicki, J. T. J. & Stachniak, Z. (eds.): *Łeśniewski's Systems Protothetic*. Nijhoff International Philosophy Series, Vol. 54. Dordrecht: Springer, 1-22.

Hugly, P. & Sayward, C. (1996): *Intensionality and Truth: An Essay on the Philosophy of A. N. Prior*. Dordrecht: Springer Science & Business Media.

CALL FOR PAPERS

We plan to dedicate a special issue of *Organon F* to modal logic. The special issue will be published in the first half of 2019 and will have the title below:

Reflecting on the legacy of C.I. Lewis: contemporary and historical perspectives on modal logic

We invite submissions presenting

- technical results in any area of modal logic, provided that they are illustrated in a detailed way and are accessible to a broad audience;
- philosophical applications (or discussing philosophical aspects) of modal logic, provided that they are connected to a relevant tradition of studies;
- historical discussions on the development of modal logic after C.I. Lewis's early works, provided that they have some relevance for contemporary investigations.

The special issue will be guest-edited by **Matteo Pascucci** (Vienna University of Technology and Slovak Academy of Sciences) and **Ádám Tamás Tuboly** (Hungarian Academy of Sciences and University of Pécs). It will include three invited articles by established researchers in the field and contributed manuscripts selected via a double-blind reviewing process. There are no restrictions regarding the style of submitted manuscripts, but the authors of accepted contributions will be asked to prepare their final versions according to the journal's guidelines. Manuscripts should be sent in **pdf** format by

15 September 2018

to any of the guest-editors (please, put ORGANON-F-MODAL-LOGIC-SUBMISSION as a subject):

matteo.pascucci@tuwien.ac.at
tubolyadamtamas@gmail.com

Feel free to contact the guest-editors for further information on the special issue and on the format of submissions.