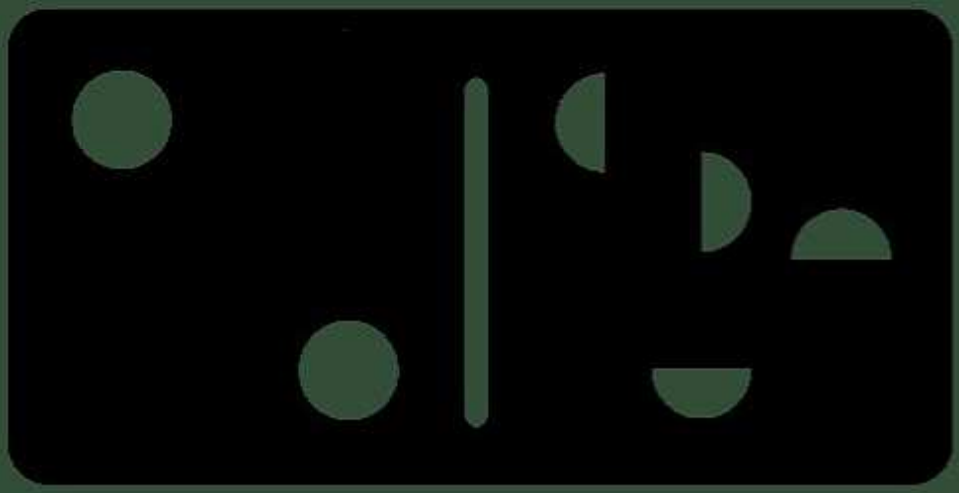
INTERNETOWY MAGAZYN FILOZOFICZNY
HYBRIS 3/2017 (38)



TRENDS IN CONTEMPORARY POLISH PHILOSOPHY OF MIND

edited by PAWEŁ GRABARCZYK
and DAWID MISZTAL

INTERNETOWY MAGAZYN FILOZOFICZNY „HYBRIS”
Instytut Filozofii Uniwersytetu Łódzkiego
Ul. Lindleya 3/5
90-131 Łódź
tel./fax: (48) (42) 635 61 35/(29)
e-mail: magazyn.internetowy.hybris@gmail.com
ISSN: 1689-4286

TRENDS IN CONTEMPORARY POLISH PHILOSOPHY OF MIND

edited by
PAWEŁ GRABARCZYK and DAWID MISZTAL

Publikacja została sfinansowana ze środków Ministerstwa Nauki i Szkolnictwa Wyższego w ramach programu Narodowego Programu Rozwoju Humanistyki przyznanych na podstawie decyzji 0014/NPRH4/H3b/83/2016 - projekt „Przygotowanie i publikacja dwóch anglojęzycznych numerów monograficznych Internetowego Magazynu Filozoficznego HYBRIS" (3bH 15 0014 83).

CONTENTS:

TRENDS IN CONTEMPORARY POLISH PHILOSOPHY OF MIND

edited by
PAWEŁ GRABARCZYK and DAWID MISZTAL

CONTENTS:

HYBRIS nr 38 (2017)
ISSN: 1689-4286

PAWEŁ GRABARCZYK
DAWID MISZTAL
UNIVERSITY OF ŁÓDŹ

INTRODUCTION: TRENDS IN CONTEMPORARY POLISH PHILOSOPHY OF MIND

The landscape of current philosophy of mind in Poland is varied and reflects most of the contemporary international trends in the subdiscipline. Its health can be easily seen by the amount of institutional backing it gets (various cognitive studies courses across the country) and the existence of specialized journals (for example *Avant* and *Studia z kognitywistyki i filozofii* umysłu). For this reason, we decide to focus mostly on one specific group of contemporary trends visible in the subdiscipline: the recent resurgence of various forms of internalism and the critical reception of this resurgence. Let us briefly outline the phenomenon in question. There are no doubts that at the end of XX century both: philosophy of mind as well as philosophy of language made a very distinct turn towards externalism. Even though classic arguments of Hilary Putnam (1975), Saul Kripke (1972), and Tyler Burge (1979) concerned linguistic (as opposed to mental)[1] content the upshot of the discussions they spawned made a great impact on philosophy of mind. One of the most visible results of this externalist tendency is the idea of embedded and extended cognition. According to the former, cognitive content is the result of an interplay between the cognitive agent and its environment. The relations between the agent and its environment are understood to be so crucial that the identity conditions of mental contents is oftentimes construed as dependent on the surroundings of the agent (similarly to how the environment was determinant for linguistic content in Putnam/Kripke's theory). The latter idea (the notion

[1] See Grabarczyk 2016 for a review of different understandings of „linguistic content" and „mental content".

of extended cognition) points out to the fact that not only the mental content but also the operations performed on this content can sometimes be delegated to external factors. What is even more important, the notion of mental content (or mental representations) started to be automatically understood in externalist terms as for several authors content has to be understood as relating to some external object, property, or event and the idea of mental content devoid of external targets seems to be almost self-contradictory (Kriegel 2008). It is in many ways fascinating to see how quickly externalism changed from a new and radical approach to the dominant perspective.

As is often the case, this dominant position resulted in a void which could then inspire a new wave of more sophisticated takes on internalism – new reasons to "turn inside". Of course, the theories characterized as such do not have to self-identify as "internalistic" or refer to internalism, but they do retain the main aspect of it: they create a space in which notions important for cognition can be defined without the appeal to external environment of the cognitive agent. Let us list some of such approaches that are relevant for the papers included in this volume and characterize them briefly.

The first notion that internalistic philosophy of mind often appeals to is the notion of computation. The reason for it is that it seems to be possible (at least in principle) to construe computation as a purely internal set of operations that is devoid of any external targets (it is, of course, still perfectly possible for elements of computations to refer to internal states of the computing machine). If it is possible to specify computations regardless of their target or application (in other words, if the identity criteria of computations do not demand us to refer to external objects), and if the notion of computation is relevant for cognition (which is, of course, a contentious claim in and of itself), then there still is some hope left for internalism in philosophy of mind. One specific subset of computations that (according to some authors) is especially relevant for cognition is the inferential subset of computations performed by the system (this inferentialist approach to cognition is especially evident in the classic example of Ned Block's theory of narrow mental content (Block 1987)).

Another important notion that creates space for contemporary internalism is the idea of structural representations championed by Cummins (1989) and revived lately in Ramsey (2007). In short, the idea

of structural representations boils down to the fact that some internal structures of the cognitive agent relate to their targets due to the fact that their structure is homomorphic to the target's structure. On the face of it, this idea is fully externalist (as the structure in question is specified by appeal to the target) but the trick is that it allows for a fully internalistic reading, because one of the characteristic aspects of structural representations is that they can be processed by the system off-line. For this reason, it is possible to imagine a situation in which a given system entertains and modifies a given representation to a point in which it loses the connection with its target (but still has some cognitive value).

The list of new internalistic ideas and trends in contemporary philosophy of mind wouldn't be complete if we did not mention the theory of predictive coding and the interrelated understanding of minds as anticipatory mechanisms (Hohwy 2013, Clark 2016). Contrary to traditional views on mind, which saw it as passive receiver and categorizer of stimuli, these theories interpret the mind as constantly constructing the reality. According to this view, the reality we live in is more similar to a simulation or conscious hallucination than to reality (understood in an old-fashioned sense). We could say that we literally live in a bubble and use the connections with our surroundings as "reality checks" – signals which help us correct and modify our predictions. Minds do not interact with reality, they live in self-constructed models of it. It is not hard to see that this radical idea (which has been in many respects anticipated by Metzinger 2009) gives hope to internalists as it relegates the role of cognitive system's environment to that of a pragmatic "checkpoint" needed only to steer our cognition in the right direction (but not to shape it).

***

Papers collected in this volume relate (directly or indirectly) to these "inward" trends of modern philosophy of mind. In a paper entitled "The false dichotomy between causal realization and semantic computation", Marcin Miłkowski shows that mechanistic understanding of computation does not prevent us from semantic considerations. As he points out (following Bechtel 2009), "computational modeling is not just about 'turning inside'. It requires looking up, down, and around". Miłkowski does not prevent computations to be devoid of meaning – on

the contrary, he admits that there definitely are examples of computation which are not semantic (in other words, the ideas of semantic content and mechanistic computation are logically independent). Still, it should be pointed out that being logically independent does not mean that there are no significant relations between mechanistic computations and semantic content. To the contrary – if present, semantic aspects constitute constraints on computation. In this sense Miłkowski shows that computational theories of cognition (specifically mechanistic ones) are in fact agnostic when it comes to the difference between internalistic and externalistic interpretations of cognition. One of the advantages of this paper is that it clearly differentiates between the social and environmental factors that could influence internal computation of a cognitive system (a difference that is well known but, sadly, often conflated). Miłkowski shows this on a very convincing example proposed by Shagrir (2006) in which the internal states of a machine can be interpreted as a conjunction or as a disjunction, depending on the social practices that surround it.

A similar line of argumentation can be found in Paweł Gładziejewski's paper "Just how conservative is conservative Predictive Processing?". Gładziejewski looks at the theory of predictive coding and shows that, contrary to what may seem to be the case at the first glance, this theory does not have to clash with the ideas of 4E cognition (embedded, embodied, extended, and enactive). Similarly to what Miłkowski does for mechanistic computationalism, Gładziejewski argues that the theory of predictive coding can be seen as agnostic in the sense that it is possible to interpret it as compatible with externalism. This idea is novel, since, as Gładziejewski points out, predictive coding "was initially construed in a manner that dovetails with traditional approaches in cognitive science, i.e. ones that see cognition as matter of inferential, exclusively intracranial processes involving richly structured representational states" (he calls it "conservative" or "radical" reading). One of the reasons why this conclusion is possible is that, as Gładziejewski points out, the notion of "inference" used in the theory of predictive coding is very liberal and differs from strict understanding proposed by Friston (2013). Specifically, the inferences proponents of predictive coding talk about should be truth preserving (which obviously ties the cognitive system with its surroundings). In addition to this, Gładziejewski shows that the methods the theory of predictive coding

uses to delineate internal and external processes do not suffice for such a demarcation. Last but not least, what the paper explains is that the type of representationalism that predictive coding appeals to has necessary ties to environment and to the surrounding social practices.

To complement the new forms of internalism (and their critical adoption), it is also good to look back at the original externalist's argumentation and reevaluate it from contemporary point of view. This task is taken by Witold Hensel in a paper entitled "Watered Down Essences and Evasive Speech Communities. Two Objections to Putnam's Twin Earth Argument". Hensel analyzes the seminal Twin-Earth thought experiment and shows that it rests on two necessary assumptions which are very hard to accept in the light of contemporary science. The first assumption is that objects referred to by a given natural-kind name contain common micro-structures (microessentialism). As pointed out by Hensel, this assumption is not corroborated by contemporary science (neither biology nor chemistry). The second, less obvious assumption Putnam makes is that it is possible to delineate different communities (and thus the intended reference of the terms they use). The problem can be presented as follows: Putnam helped us realize that all natural-kind terms have a hidden indexical component that ties them to a given environment. For example – the term "water" used by inhabitants of Earth was always used as referring to the microstructure of a specific liquid found on Earth. But why should we treat linguistic communities of Earth and Twin-Earth as separate? It is not obvious why should the boundary be put in this particular place, but this ability seems to be presumed in Putnam's argumentation.

An interesting illustrations of the tension between external and internal perspective can also be seen in the papers of other authors. Katarzyna Kobos discusses the situations when perception occurs in the absence of sensory stimuli. Can we say we actually perceive anything in such circumstances? Can silence be said to be heard or darkness to be seen? What is the role of the brain (if any) in forming of sensory experience? To what extent the latter is dependent on external input? In her attempt to answer these questions, Kobos meticulously analyzes two models of perceptual response to the absence of sensory impingement. Consequently, she turns to embodied predictionism as it seems to be more theoretically satisfying and more promising in terms of its explanatory power.

Marek Prokopski brings our attention to the problem of other minds. The author is mainly interested in the conceptual formulation of said problem (as opposed to ontological and epistemological formulations) which poses the question of possibility of universal mental concepts describing emotional states or inner experience. In other words, can we – asks Prokopski – justifiably use the mental concept of – say – pain, based on personal experience of pain, not only in the first person but also in third person cases? The challenge here, according to the author, is "to develop plausible positive account of mental concepts", since the negative one would lead to the disputable conclusion that we have two different mental dictionaries: a first-personal and a third-personal.

As it is often the case with opposing theoretical proposals, however incompatible they may seem, they may be inspirational for searching possible ways to reconcile them. Przemysław Nowakowski's interesting attempt to integrate computational and embodied approach to cognition can be read precisely in this context. However, to achieve his goal the author adopts an interalist rather than externalist perspective on the evolution of cognition. He assumes that internal complexity of organism is at least equally important in evolutionary shaping of cognitive processes as external, environmental factors. On this basis, Nowakowski presents his own approach to embodied cognition which he dubs E-codes' approach (E-codes being "Efficient, robust and body-specific processing"). And what he hopes to obtain by means of this approach is to create an opportunity for developing conceptualizations that would do justice not only to the embodiment thesis but to empirical data as well. Although, as he cautiously remarks, that would require "more comparative meta-analysis and computational modeling than psychological experiments".

The debate between internalism and externalism is continued in the next two chapters. In the first one by Krystyna Bielecka, this opposition is thoroughly examined in the context of the problem of intentionality, and the focus is on the semantic internalism as a potential solution to this problem. Analyzing the notion of narrow content (which basically means a content limited to its functional role within the cognitive system) in its radical interpretations, the author presents detailed critique of the aforementioned stance. In her opinion, semantic internalism deprives the content of any other than formal (i.e. syntactic)

properties, and thus it renders ascribing truth to representations (or any other semantic property for that matter) impossible.

The anti-internalistic tone of Bielecka's text is seemingly further reinforced in the last section of our book in which Maria Matuszkiewicz offers her exhaustive discussion of Robert Stalnaker's work entitled *Our Knowledge of the Internal World* (2008). The chapter identifies and elaborates the central issues of Stalnaker's argument such as our epistemic relation to our experience, the relation between experience and knowledge, or the relation between objective knowledge and the knowledge we can have only from a certain perspective. But Matuszkiewicz not only fully exposes Stalnaker's version of externalism, pointing additionally to its affinity with other philosophical positions (with contextualism, for example). She also notices that Stalnaker's externalism, being rather a methodological perspective than metaphysical view, is not altogether so anti-internalistic as it may seem at first glance.

REFERENCES

Bechtel, W. 2009. Looking Down, Around, and up: Mechanistic Explanation in Psychology. *Philosophical Psychology* 22 (5): 543–64. doi:10.1080/09515080903238948.

Block, N. (1987), Advertisement for a Semantics for Psychology. *Midwest studies in philosophy* 10.1, p. 615-678.

Burge, T., 1979. Individualism and the Mental, in French, Uehling, and Wettstein (eds.) *Midwest Studies in Philosophy*, IV, Minneapolis: University of Minnesota Press, pp. 73–121.

Clark, A. (2016b). *Surfing Uncertainty. Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

Cummins R., (1989), *Meaning and Mental Representation*, Cambridge, MA, MIT Press.

Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, 10, 20130475–20130475.

Grabarczyk, P., 2016, How Meaning Became "Narrow Content", *Studies in Logic, Grammar and Rhetoric*, vol 46, issue 1.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Kriegel, U. (2008). Real Narrow Content. *Mind and Language*, (23), 305–328.

Kripke, S., 1972. *Naming and Necessity*, Oxford: Blackwell.

Metzinger, T. 2009, *The Ego Tunnel*, New York: Basic Books

Putnam, H., 1975. The Meaning of Meaning, *Philosophical Papers, Vol. II: Mind, Language, and Reality*, Cambridge: Cambridge University Press.

Ramsey W.M., (2007), *Representations Reconsidered*, Cambridge, MA, Cambridge University Press.

Shagrir, O., 2006. Why We View the Brain as a Computer. *Synthese* 153 (3): 393–416. doi:10.1007/s11229-006-9099-8.

HYBRIS nr 38 (2017)
ISSN: 1689-4286

MARCIN MIŁKOWSKI
INSTITUTE OF PHILOSOPHY AND SOCIOLOGY
POLISH ACADEMY OF SCIENCES

THE FALSE DICHOTOMY BETWEEN CAUSAL REALIZATION AND SEMANTIC COMPUTATION

It has been argued that there is a tension between the semantic characterization of neural computation and the causal account of computation (Shagrir 2010). Just because the heuristic role of semantic entities in cognitive science is difficult to deny (Bechtel 2016), it might be thought that that the causal account is descriptively inadequate for our current scientific practices. Moreover, others have claimed that there is a role for content-involving computation (Rescorla 2013) in computational explanations. If there is, it means that the mechanistic account of computational explanations misses an essential aspect of the scientific practice.

However, I will argue that semantic computation and the causal account of neural computation are not mutually exclusive, and they both have important explanatory, descriptive, and heuristic roles. One does not have to decide to embrace the mechanistic account on pain of rejecting all semantic considerations; this is a false dichotomy. In particular, semantic notions usually require rich interactions with the environment and appropriate internal orchestration of the mechanism; purely computational modeling is usually limited to the internal functioning of a mechanism, while there are complex inter-level and intra-level relationships between computational, semantic, and, more broadly speaking, causal posits in explanatory models in neuroscience.

In this paper, I will show how semantic factors constrain the understanding of the phenomena to be explained so that they naturally help build better mechanistic models. In section 1, I will elucidate why

one could think that there is a tension between mechanistic accounts of physical computation and semantic computation in general. Next, in Section 2, it will be argued that understanding of what cognitive systems may refer to is important in building better models of their cognitive processes by specifying the function of cognitive mechanisms *partially* in content-involving ways. For this purpose, a recent study of some phenomena in rats that are capable of 'entertaining' future paths (Pfeiffer and Foster 2013) will be analyzed in Section 3. The researchers stress that the hippocampus 'generates brief sequences encoding spatial trajectories', which is a clearly semantic way of framing the phenomenon. The above case shows that computational modeling is not just about 'turning inside'. It requires looking up, down, and around (Bechtel 2009). Looking around requires one to understand the environmental structure. In short, computation and representation, considered in an externalist fashion, do not screen off each other. Why should they? Representing requires physical information, and functional physical information processing amounts to physical computing.

1. The tension between causal realization and semantic computation

The purpose of mechanistic accounts of physical computation is to deliver a normatively and descriptively adequate list of necessary and sufficient conditions that physical systems must satisfy to qualify as computers. There are some differences between these accounts (Miłkowski 2013; Piccinini 2015), yet they may be summarized jointly in the following way. The necessary condition for candidate physical systems is that they be mechanisms (in the sense of the new mechanistic philosophy, cf. (Machamer, Darden, and Craver 2000; Bechtel 2008; Craver 2007)) whose function is to compute. The mechanism's causal structure should correspond strictly to a mathematical model of computation over physical vehicles specified in a substrate-neutral way. Moreover, the computational explanation should essentially involve processing of information (as Miłkowski states the condition) or be usable as information (as Piccinini has framed it). The rest of conditions spelled out by Miłkowski and Piccinini simply follow from the general methodological norms of mechanistic explanation.

One striking feature of the mechanistic account is that it does not require vehicles of computation to be semantic in any rich sense. In other

words, mechanists explicitly reject the claim that only physical systems whose parts are semantic can be computers (Piccinini 2008; cf. Fresco 2010). They assume that there may be computers that operate on symbols without any denotation or intrinsic meaning. But this is not because they share the conviction that semantic notions are disposable altogether. Rather, they think that semantic notions are more difficult to specify than the conditions of physical computation. David Chalmers has long argued in the same vein:

> If we build semantic considerations into the conditions for implementation, any role that computation can play in providing a foundation for AI and cognitive science will be endangered, as the notion of semantic content is so ill-understood that it desperately needs a foundation itself (Chalmers 2011, 336).

As such, mechanistic and causal accounts refrain from semantic considerations. For this reason, however, they can be criticized. First of all, there is an important role of cognitive representations in cognitive explanations. For example, the whole history of research on the cognitive maps in rats was based on a strong assumption that they refer in various ways to their environment, and it has resulted in a very promising research program (Bechtel 2016). But this role seems to be irrelevant to the mechanistic account.

Second, it has been argued that mechanists cast their net too wide which results in limited pancomputationalism: they would have to admit, as Chalmers does, that a rock implements a trivial computation – or even worse, a class of trivial computations specified as *any* constant function. Namely, the rock's position may be considered to encode the result of the computation. Of course, the rock does not implement *all possible* computational functions, but still a lot of them (Shagrir 2006, 398, 2010, 272). But Miłkowski (2013, 79), for example, denies that a rock is a computer: a computational explanation of the rock's behavior is not any more predictive nor has any more explanatory power than a physical one in terms of gravity, which explains why the rock does not fly away etc. Furthermore, the rock's function is not to compute; no parts of the rock were selected according to any design as types to perform the constant functions (Miłkowski 2013, 62). Piccinini also requires that the result of the computation be usable: "the important point is that we are interested in computation because of what we (finite observers) can learn from it"

(Piccinini 2015, 256). So, while it could still be argued that the semantic constraints do not restrict the class of the candidate physical computers, other constraints allow mechanists to avoid the charge of drawing the boundary between computational and non-computational systems in a wrong way.

A third objection is much more difficult to handle *prima facie* (cf. Shagrir 2006, 409; the example has been simplified). Imagine two electrical circuits, CIRC1 and CIRC2. The first responds with voltage v_2 whenever it receives v_2 and v_2 on its input, otherwise it responds with v_1; whereas the second responds with v_1 whenever it receives voltage v_1 and v_1 on input, and otherwise with v_2. Which one of these is the OR gate that corresponds to inclusive disjunction, and which is the AND gate, the device for computing conjunction? If we treat v_1 as true, and v_2 as false, then CIRC1 is an OR gate, and CIRC2 implements an AND gate. But we might switch the logical interpretation, and then CIRC1 is an AND gate, and CIRC2 an OR gate. In other words, it seems that there are two empirically adequate but inconsistent mechanistic explanations of CIRC1 and CIRC2. This would mean that the mechanistic account is deficient and clearly worse than the semantic account. The semantic account, after all, can constrain the interpretation of voltages by taking into account the use of the circuit in its environment and possibly in a larger computational context.

Note, however, that if we have no further information about how the circuit is used, the semantic account fares no better. There is no fact of the matter that could restrict possible interpretations. So what kind of information could restrict explanations in this case? For example, there could be also one-input circuits that respond with v_2 to v_1, and vice versa. These are probably NOT gates, but we still have no way to say how to assign truth and false to voltages. But there are frequent combinations of NOT gates and CIRC1 gates. As this combination in a disjunctive normal form for propositional calculus corresponds to a material implication realized as NOT + OR, we could settle for the interpretation of v_1 as true, and v_2 as false. This is a purely syntactic hypothesis. We could also see that a device responds to two input data (for example, from its receptor devices) by using CIRC2 gate, and then v_2 triggers some response. A semantic hypothesis could be that these inputs need to be both present for the whole system to respond; so the system uses a conjunction of two receptor values. This is again a semantic hypothesis, which seems to

confirm the first one. But it's definitely not sufficient in itself, as it does not allow us to reject the hypothesis that the receptors are actually silent and that what one sees is the false disjunction. In short, it takes a lot of experimentation and careful consideration to decide such issues (and it may be impossible to decide which logical connectives are at play as based merely on stimuli and responses also in the human case, cf. (Berger 1980)). It does not seem, therefore, that one account fares better than another in this case; the case is indeed difficult. However, it may motivate the claim that the mechanistic account should not restrict itself to purely formal considerations. How the mechanism responds to the environment may be essential for explaining it.

The difficult case above is similar to the one sketched in the argument put forward by Michael Rescorla (2013, 686). While Rescorla does not endorse the semantic view on computation, he claims that there are content-involving instructions in computer programs. This claim is defended against all structuralist accounts of physical computation, not only against the mechanistic view. Content-involving instructions depend in their causal efficacy on the wide social context of the use of computers; an example of this may be the dependence of the numerical notation of numbers in a programming language Scheme. It is executed on two machines in two different societies: one uses base-10 notation, and another base-13 notation, so the program to compute the greatest common divisor of 115 and 20:

```
(gcd 115 20)
```

correctly yields '5' in the base-10 society, but incorrectly in the base-13 society because '5' "is not a divisor of the base-13 denotation of '20' (namely, the number twenty-six)" (Rescorla 2013, 688).

However, the example does not fully prove the point. The problem is that the type of numerical notation is explicitly defined *syntactically* in Scheme. Specifically, it is defined in Backus-Naur Form (BNF), which is a syntactical tool used (usually with numerous extensions) to define programming languages. The format numbers is defined in the section 4.2.8, which is a part of Chapter 4 "Lexical syntax and datum syntax" of the official language specification (Flatt et al. 2009). Here are the definitions of decimal digits and hexadecimal digits:

<digit> → 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

<hex digit> → <digit> | a | A | b | B | c | C | d | D | e | E | f | F

It would be difficult to encode base-13 using <digit> as defined above, as one would not be able to write out A, B, or C in 13-base notation (which correspond to 10, 11, and or 12 in the decimal notation). There are missing symbols, at least according to standard encoding conventions used in programming (note: one could have a non-standard notation that would treat one of the digits as special, and not use a simplistic positional encoding). In other words, while the base-13 society wrongly thinks that Scheme assumes the base-13 notation, it makes no difference as to what program is physically implemented. If there are any facts about programming languages such as Scheme, the base-13 society got them wrong.

To see that they could be shown to be wrong, it is useful to remind how language compilers or interpreters are evaluated. A series of tests, called *regression tests*, are devised in a given programming language. The execution of such tests triggers a number of assertions embedded in the test. For example, one can assert that (gcd 115 20) yields '5'. A failure of the assertion means that the compiler does not conform to the language specification.

Similarly, the fact that a user thinks that Microsoft PowerPoint is a word processing program it does not make PowerPoint a word processing program. The user is simply wrong. Of course, it might be objected that if a society had used PowerPoint for its word-processing needs, PowerPoint *would become* a word-processing application. In other words, the intention of the software application developers may not determine the function of the application, just like the intentions of designers of technological artifacts do not fully determine their functions (amulets do not really have their functions). While the issue of technical functions of artifacts is vexed, the general consensus is that one of the determining factors of technical function is also the users' intention, rather than the designer intention (see, e.g. Vermaas and Houkes 2006).

While the simple numerical notation example introduced by Rescorla does not satisfactorily show that the mechanistic approach is deficient, there is a deeper point there. The point can be easily proven by adapting the example and using some notation that would use, say 8-base, as there would be no symbols missing for the BNF specification (Rescorla, personal communication). In such a situation, it would be

impossible to determine the interpretation of '100': one society would understand it to stand for decimal 64, and another for decimal 100. And there is, potentially, an infinite number of similar ambiguities inherent in programming languages.

To sum up, the function of mechanisms may depend on their social and widespread use, and the use may involve semantic factors. People frequently use computers to manipulate their external representations. Indeed, the rest of the paper will argue that the proper focus on the function of mechanisms shows that semantic considerations may play a serious role in computational explanations considered mechanistically. Mechanistic explanation should, at least for an important class of computational mechanisms, include semantic considerations.

2. Building mechanistic models by including semantic constraints

In this section, the notion of function used in mechanistic accounts of computation will be made more explicit. Then it will be shown that some but not all computational mechanisms have semantic functions (in a sense to be elucidated below). These functions will be only partially explained computationally. However, they will constrain the space of plausible computational mechanisms posited in mechanistic explanations.

Mechanistic accounts of physical computation focus predominantly on functional mechanisms (cf. Garson 2013). However, there is a debate over the notion of function appropriate for mechanistic explanations. Most defenders of mechanistic explanations rely on a fairly weak account that equates function with a capacity of a given physical system — its capacity to perform some causal role owning to its internal organization (Cummins 1975)— that is of epistemic interest (cf. Craver 2013, 2001). While it is a fact of the matter whether the system has such function or not, the ascription is based on the perspective taken by a beholder. But defenders of the mechanistic account of computation do *not* embrace the perspectivalist view: they argue that the mechanistic account of physical computation should avoid, if possible, any appeal to epistemic interest of beholders, since numerous objections against the possibility of an objective account of physical computation rely on the possibility of arbitrary ascriptions of computations to physical systems. Moreover, they want to account for malfunction of computational

systems or the failure of physical mechanisms to perform their function. One of the major objections against the perspectivalist view is that the same capacity of a physical system may count as functional and dysfunctional at the same time (Millikan 2002). Instead, Piccinini and Miłkowski have both argued for teleological accounts. While there are some notable differences between their accounts, they both seem to embrace a unified view on a function that includes technical functions of artifacts and teleological functions of natural computing mechanisms. For example, Piccinini defines the notion in the following way:

> A teleological function (generalized) is a stable contribution to a goal (either objective or subjective) of organisms by either a trait or an artifact of the organisms (Piccinini 2015, 116).

The upshot of this definition is that there cannot be any computers without organisms: either as their users or as physical mechanisms whose goals are satisfied by the existence of such computers. Quite clearly, before there were organisms, there were rocks, and they were not implementing any functions. So far, so good. But couldn't it be possible in principle that there could exist computational physical systems other than organisms or artifacts produced by organisms? For example, one could imagine naturally evolved robots that have their goals fulfilled thanks to computation. But Piccinini rejects this possibility by saying that these would count as organisms in a broader sense (Piccinini 2015, 113).

The approach of Miłkowski is partially similar to the one proposed by Craver and Cummins but also relies on the teleological view defended at length by Ulrich Krohs (2004, 2007): "the functional role of a component is one of its causal roles, such that it contributes to the system behavior of the mechanism (as in the classical analytical account in Cummins 1975; for a mechanistic variant of this account, see Craver 2001), but the organization of the mechanism is based on the process of selection of its parts as types" (Miłkowski 2013, 62). This requires a bit more elucidation. Krohs defends a design-based notion of function where design is understood as a type fixation of a complex entity. The type-fixed entity is defined thus:

> (COM) A complex entity is type-fixed iff its components are type-fixed.

> (TF) A component of an entity is type-fixed iff it is part of the entity because of its type and not merely because of its properties (Krohs 2007, 77).

Again, the components of rocks are not selected as types: there is no assembly process that generates them for the purpose of computing constant functions. Yet, in contrast to Piccinini, no appeal is made to the existence of organisms.

It's beyond the scope of this paper to compare both accounts in detail, and see how they address the main objections in the debate over teleological function. Still, it's instructive to discuss shortly an alternative view on technical functions. For example, a sophisticated ICE theory (Intentional-Causal role-Evolutionist) is defended by Pieter Vermaas and Wybo Houkes:

> An agent *a* ascribes the capacity to ϕ as a function to an artefact *x*, relative to a use plan *p* for *x* and relative to an account *A*, iff:
>
> I. the agent *a* has the capacity belief that x has the capacity to ϕ, when manipulated in the execution of *p*, and the agent *a* has the contribution belief that if this execution of *p* leads successfully to its goals, this success is due, in part, to *x*'s capacity to ϕ;
>
> C. the agent *a* can justify these two beliefs on the basis of *A*; and
>
> E. the agents *d* who developed *p* have intentionally selected *x* for the capacity to ϕ and have intentionally communicated *p* to other agents *u* (Vermaas and Houkes 2006, 9).

Note that this account rules out ascriptions of computational functions to biological brains, as they were not selected by any intelligent agent.[1] However, it can be easily used to ascribe functions to a computer running a Scheme interpreter or to a pair of logical gates. One can consult the agents who have developed the Scheme interpreter and determine that base-13 society is indeed wrong in assuming that '5' is given in base-13 notation (see section 1). In other words, under ICE account, semantic considerations may be framed in terms of the developers' intentions. And

[1] At least most of them, except for direct genetic modifications, such as the ones used in optogenetics (Deisseroth et al. 2006)

these considerations may constrain the hypotheses about the function of computational artifacts.

A similar move is possible under Miłkowski's account, as long as the type fixation process is sensitive to semantic values of computations performed. For example, one may analyze the compiler or interpreter of Scheme programming language to see whether the results of defined numerical functions turn out to be systematically correct and coincide with the BNF specification. The BNF specification, after all, was most probably used to design the compiler or interpreter (it makes no difference to this account whether it was this particular specification or some other). And the same can be done using Piccinini's account: the goals of organisms using Scheme on their computers will be achieved if the Scheme interpreter or compiler is executed, so the computer may be ascribed a function to run Scheme programs (interpreted or compiled), and thus to execute any function the user might want to execute. So, while mechanistic accounts of function are more general, in terms of semantic considerations, they do not fall behind sophisticated accounts of technical functions.

The upshot of this short discussion is that the gist of considerations cited in favor of semantic accounts of computation can be preserved in the mechanistic account. For example, Jerry Fodor has claimed that it's characteristic for (some) mental processes to preserve semantic properties such as truth. In his opinion, what makes computational psychology so compelling is the fact that one may build a computer that does the same:

> if you have a device whose operations are transformations of symbols, and whose state changes are driven by the syntactic properties of the symbols that it transforms, it is possible to arrange things so that, in a pretty striking variety of cases, the device reliably transforms true input symbols into output symbols that are also true. I don't know of any other remotely serious proposal for a mechanism that would explain how the processes that implement psychological laws could reliably preserve truth (Fodor 1995, 9).

While mechanists have pointed out that there could be computational processes that do not preserve the constraint of truth preservation — a trivial counterexample is a single NOT gate – there are plenty that do. So while preservation of semantic properties is not an essential property of

computational mechanisms, it is a property that can be partially explained computationally in terms of reliable processes of computation over vehicles that were arranged in a manner that preserves semantic constraints. Simply, one cannot explain truth preservation unless there are also appropriate syntactic processes. This is what can be explained computationally about representation; so even if intentionality cannot be reduced to computation, some regularities in intentional processes can be explained computationally.

In semantic computation, the vehicles over which the computations are performed are bearers of semantic information. Notice that a vehicle *cannot* have semantic properties if it is not a bearer of structural information (data): the data needs to be well-formed to have semantic content. The condition of well-formedness of data is always satisfied for computational mechanisms according to the mechanistic account of physical computation. But computational mechanisms need not operate on meaningful data. They may as well process gibberish.

In general, two kinds of semantic information may be distinguished: instructional and factual (Floridi 2010, 34). The first conveys the need for a specific action, and the latter states the facts. While it is not controversial that in programmable computers there are programs full of instructional information (Fresco and Wolf 2013) it is far from obvious that one can build computers whose symbols are genuinely or intrinsically meaningful in the factual sense (Harnad 1990). The mechanistic account of physical computation does not presuppose, therefore, that all computation is over meaningful data. However, it does not exclude the possibility of computation over meaningful data. In this, it clearly differs from the semantic view defended by Shagrir, and at the same time, it can include semantic constraints in mechanistic explanations. This also means that the mechanistic account is not *merely* structural: it may appeal to content-involving facts, such as the ones invoked by Rescorla.

While the account of what makes well-formed data semantic goes beyond the scope of this paper (but see Floridi 2010; Dretske 1982; MacKay 1969), there are mechanistic explanations of representational phenomena. Mechanists presuppose that intentionality or semantic properties may be explained in terms of semantic information and teleological function, and some have already proposed accounts of representational or intentional mechanisms (Miłkowski 2015; Plebe and

De La Cruz 2016). Representational mechanisms are an important proper subset of computational mechanisms.

The assumption that a given mechanism is representational constrains computational hypotheses about the system; here, the mechanistic account follows Shagrir's (2001) analysis. Let's take the example of ambiguous circuits, CIRC1 and CIRC2. If we know how these circuits are supposed to work – what their representational function is, i.e., what kind of characteristics of entities are represented by computational vehicles – we can settle for one interpretation of the voltages in the circuits. To wit, the mechanistic account, thanks to the notion of the representational function of computational mechanisms, can make use of the considerations cited by Shagrir and Rescorla. In the next section, one case will be studied in detail to show how.

3. Semantic constraints at work

Cognitive maps are paradigmatic examples of genuine mental representations cited by neuroscience. The representational hypothesis, put forward by Edward Tolman (1948), has inspired a particularly rich research program (Bechtel 2016). Such maps are structured but not reducible to language-like media (Rescorla 2009); they are also prime examples of structural representations (Cummins 1996). While there are multiple different mechanisms involved in the functioning of cognitive maps – different kinds of cells are responsible for representing distinct features of the environment in quite complex ways, a recent finding of representing future paths as trajectories to a goal will be analyzed here. The finding concerns a neural code discovered in the rat's hippocampus.

The rat's hippocampus generates brief sequences encoding spatial trajectories strongly biased to progress from the subject's current location to a known goal location. Pfeiffer and Forster (2013) were able to find direct evidence for the existence of future-focused navigational activity of place cells in a realistic two-dimensional environment. They have elegantly shown that it is related to sharp-wave-ripple (SWR) events; SWRs are irregular bursts of brief (100–200 ms) large-amplitude and high-frequency (140–200 Hz) neuronal activity in the hippocampus. In other words, there is direct evidence that place cells are involved in planning future routes. To find this evidence, Pfeiffer and Foster used a 40-tetrode microdrive that permitted synchronous electrophysiological

activity recording from 250 place cells. Using sophisticated mathematical methods, they were able to decode the locations represented by this cell ensemble in SWRs.

However, the finding is all the more exciting because it can be integrated with previous work on cognitive maps (Schmidt and Redish 2013). This previous work is also computational. A number of computer simulation studies were designed to study cognitive maps and their possible neural encodings (see e.g. McNaughton et al. 2006; Conklin and Eliasmith 2005). Simulations take inspiration from experimental results and often go beyond available evidence, and experiments are then designed to test for plausible computational schemes. Neuroscientists understand that there are neural structures that have special computational roles, but that doesn't mean that a single anatomical structure plays just one role; as it turns out, it may play multiple roles in multiple neural systems, which is evidenced in the work on the hippocampus (Redish 1999, xiii). The neural code used to plan future routes is yet another code among the ones already discovered in navigation computations performed by the rat.

From the mechanistic point of view, current computational models, impressive as they are, remain incomplete because of the intrinsic complexity of the navigational subsystems and difficulties involved in their study. What is notable here is that Pfeiffer and Foster assume a representational point of view and explore the electrophysiological activity of neurons as related to the features of the external environment in the rat subject in various experimental conditions. In the discussed experiment, rats foraged for food distributed in random locations. Every day, they would start from the same home location, which remained constant for the day, and would change the next day. This way, rats could try novel routes. In other experiments, rats may learn the topology of the maze and then they are transferred to similar mazes to discover how they remember the topology (Alme et al. 2014). In other words, what is studied is the relationship between the activity of the organism and its environment. Only in such a context does a computational model make sense; and the overarching hypothesis is that neural processes are involved in various representational tasks.

The discovery of encoding requires researchers to understand what features of the environment could be encoded by neural events, and then to study (statistically) the results of electrophysiological recordings

as related to these features. In the study under analysis, the researchers have found that there are two kinds of trajectory events: ones that were initiated when the rat was at the Home location ('home events'), and the ones initiated elsewhere ('away events'). Interestingly, it turns out that the Home location was over-represented in away-events relative to other locations in the open field. This means that researchers need not presuppose that representation in the brain is absolutely veridical; it may be biased for some reason (one may speculate, for example, that the Home location is particularly important because the rat started its exploration there). So how can they be sure that these trajectory events really represent future routes? The confirmation of this representational hypothesis is that the rat simply takes one of the future routes immediately after planning it.

The trajectory events discovered by Pfeiffer and Foster are consistent with the number of previous hypotheses and allow researchers to make them more precise by offering an experimental method:

> trajectory events relate to hippocampal function in multiple conceptual contexts: as a cognitive map in which routes to goals might be explored flexibly before behaviour, as an episodic memory system engaging in what has been termed 'mental time travel', and as a substrate for the recall of imaginary events. These conceptualizations reflect a continuity with earlier speculations on animals' capacities for inference (Pfeiffer and Foster 2013, 78).

In other words, understanding the context in which a given mechanism works helps the modelers to analyze its internal structure that is supposed to perform inferential computations, especially those related to mental time travel, route planning, and the recall of imaginary events. The experimental method yields semantic constraints on computational models of these inferential processes: plausible models should conform to neural encoding schemes discovered experimentally. Otherwise, computational models of the hippocampus might diverge from what is known about the behavioral functioning of the rat, and this is precisely what researchers want to avoid. In terms of the mechanistic approach to explanation, one may state it in the following way: The phenomenon to be explained is described as the function of place cells to represent future paths, and the causal explanation (currently somewhat incomplete, as

precipitating conditions of the mechanism are not clear) shows the orchestrated activity of place cells that contributes to the realization of this function.

It needs to be noted that computational models are in general difficult to confirm or disconfirm experimentally; one may usually produce a number of *different* models consistent with experimental findings. Including more constraints allows researchers to reject at least some models. This way modeling is less arbitrary. In some sense, modelers need to practically solve the ambiguities such as the ones mentioned by Shagrir in his example of experimentally ambiguous logical gates, or by Rescorla in his example of ambiguous numerical encoding. They may do it by including semantic constraints in the specification of the explanandum phenomenon.

To sum up, it is only natural to assume that the function of neural mechanisms involved in solving representational tasks is to represent. There is no particular reason to abstain from representational hypotheses, which are extremely helpful from the mechanistic perspective to make models explanatorily more plausible.

4. Conclusion

Successful cognitive modeling is a question of satisfying multiple constraints from multiple fields of inquiry, levels of organization, and theories. Semantic and ecological considerations are not just heuristics of discovery of mechanisms. They are constraints over the space of possible mechanism representations. By a *constraint* I understand a representation that shapes the boundaries of the space of plausible representations of mechanisms or the probability distribution over that space (Miłkowski 2017). The more constraints are satisfied, the more integrated the model of a mechanism becomes. Ideally, all constraints should be satisfied to produce an explanatorily plausible mechanism model.

The mechanistic view on physical computation does not assume that all computation makes sense. There may be plenty of computation without any representational role. However, there are computations over representations, and these are extremely important for cognitive (neuro)science. For this reason, to remain descriptively and normatively adequate, the mechanistic view has to assume that representational constraints are important, and they can be naturally included in

descriptions of functions of computational / representational mechanisms.

Hence, the dichotomy between the causal realization and semantic computation is false. Semantic computations are realized causally, and they can be studied mechanistically. For the mechanistic account of explanation, there is no reason to abstain from representational hypotheses in science. The proponents of the mechanistic account of physical computation only stress that not all computers operate on semantic information. But computation and representation do not screen off each other.

REFERENCES

Alme, Charlotte B., Chenglin Miao, Karel Jezek, Alessandro Treves, Edvard I. Moser, and May-Britt Moser. 2014. "Place Cells in the Hippocampus: Eleven Maps for Eleven Rooms." *Proceedings of the National Academy of Sciences* 111 (52):18428–35. https://doi.org/10.1073/pnas.1421056111.

Bechtel, William. 2008. *Mental Mechanisms*. New York: Routledge (Taylor & Francis Group).

———. 2009. "Looking Down, Around, and up: Mechanistic Explanation in Psychology." *Philosophical Psychology* 22 (5):543–64. https://doi.org/10.1080/09515080903238948.

———. 2016. "Investigating Neural Representations: The Tale of Place Cells." *Synthese* 193 (5):1287–1321. https://doi.org/10.1007/s11229-014-0480-8.

Berger, Alan. 1980. "Quine on 'Alternative Logics' and Verdict Tables." *The Journal of Philosophy* 77 (5):259–77. https://doi.org/10.2307/2025755.

Chalmers, David J. 2011. "A Computational Foundation for the Study of Cognition." *Journal of Cognitive Science*, no. 12:325–59.

Conklin, John, and Chris Eliasmith. 2005. "A Controlled Attractor Network Model of Path Integration in the Rat." *Journal of Computational Neuroscience* 18 (2):183–203. https://doi.org/10.1007/s10827-005-6558-z.

Craver, Carl F. 2001. "Role Functions, Mechanisms, and Hierarchy." *Philosophy of Science* 68 (1):53–74.

———. 2007. *Explaining the Brain. Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

———. 2013. "Functions and Mechanisms: A Perspectivalist View." In *Functions: Selection and Mechanisms*, edited by Philippe Hunemann, 133–58. Dordrecht: Springer.

Cummins, Robert. 1975. "Functional Analysis." *The Journal of Philosophy* 72 (20):741–65.

———. 1996. *Representations, Targets, and Attitudes*. Cambridge, Mass.: MIT Press.

Deisseroth, Karl, Guoping Feng, Ania K Majewska, Gero Miesenböck, Alice Ting, and Mark J Schnitzer. 2006. "Next-Generation Optical

Technologies for Illuminating Genetically Targeted Brain Circuits." *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience* 26 (41):10380–86. https://doi.org/10.1523/JNEUROSCI.3863-06.2006.

Dretske, Fred I. 1982. *Knowledge and the Flow of Information*. 2nded. Cambridge, Mass.: MIT Press.

Flatt, Matthew, Anton van Straaten, Robby Findler, and Jacob Matthews. 2009. *Revised[6] Report on the Algorithmic Language Scheme*. Edited by Michael Sperber. Cambridge; New York: Cambridge University Press. http://www.r6rs.org.

Floridi, Luciano. 2010. *Information: A Very Short Introduction*. Oxford: Oxford University Press.

Fodor, Jerry A. 1995. *The Elm and the Expert: Mentalese and Its Semantics*. Cambridge, Mass.: MIT Press.

Fresco, Nir. 2010. "Explaining Computation Without Semantics: Keeping It Simple." *Minds and Machines* 20 (2):165–81. https://doi.org/10.1007/s11023-010-9199-6.

Fresco, Nir, and Marty J. Wolf. 2013. "The Instructional Information Processing Account of Digital Computation." *Synthese* 191 (7):1469–92. https://doi.org/10.1007/s11229-013-0338-5.

Garson, Justin. 2013. "The Functional Sense of Mechanism." *Philosophy of Science* 80 (3):317–33. https://doi.org/10.1086/671173.

Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D* 42:335–46.

Krohs, Ulrich. 2004. "Der Begriff Des Designs." In *Eine Theorie Biologischer Theorien. Status Und Gehalt von Funktionsaussagen Und Informationstheoretischen Modellen*, 59:70–119. Berlin: Springer.

———. 2009. "Functions as Based on a Concept of General Design." *Synthese* 166 (1):69–89. https://doi.org/10.1007/s11229-007-9258-6.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. "Thinking about Mechanisms." *Philosophy of Science* 67 (1):1–25.

MacKay, Donald MacCrimmon. 1969. *Information, Mechanism and Meaning*. Cambridge: M.I.T. Press.

McNaughton, Bruce L, Francesco P Battaglia, Ole Jensen, Edvard I Moser, and May-Britt Moser. 2006. "Path Integration and the Neural

Basis of the 'Cognitive Map'." *Nature Reviews. Neuroscience* 7 (8):663–78. https://doi.org/10.1038/nrn1932.

Miłkowski, Marcin. 2013. *Explaining the Computational Mind*. Cambridge, Mass.: MIT Press.

———. 2015. "Satisfaction Conditions in Anticipatory Mechanisms." *Biology & Philosophy* 30 (5):709–28. https://doi.org/10.1007/s10539-015-9481-3.

———. 2016. "Unification Strategies in Cognitive Science." *Studies in Logic, Grammar and Rhetoric* 48 (1):13–33. https://doi.org/10.1515/slgr-2016-0053.

Millikan, Ruth Garrett. 2002. "Biofunctions: Two Paradigms." In *Functions: New Essays in the Philosophy of Psychology and Biology*, edited by Andrew Ariew, Robert Cummins, and Mark Perlman. New York: Oxford University Press, USA.

Pfeiffer, Brad E, and David J Foster. 2013. "Hippocampal Place-Cell Sequences Depict Future Paths to Remembered Goals." *Nature* 497 (7447). Nature Publishing Group:74–79. https://doi.org/10.1038/nature12112.

Piccinini, Gualtiero. 2008. "Computation without Representation." *Philosophical Studies* 137 (2):205–41. https://doi.org/10.1007/s11098-005-5385-4.

———. 2015. *Physical Computation: A Mechanistic Account*. Oxford: Oxford University Press.

Plebe, Alessio, and Vivian M. De La Cruz. 2016. *Neurosemantics*. Vol. 10. Studies in Brain and Mind. Cham: Springer International Publishing. http://link.springer.com/10.1007/978-3-319-28552-8.

Redish, A. David. 1999. *Beyond the Cognitive Map: From Place Cells to Episodic Memory*. Cambridge, Mass.: The MIT Press.

Rescorla, Michael. 2009. "Cognitive Maps and the Language of Thought." *The British Journal for the Philosophy of Science* 60 (2):377–407. https://doi.org/10.1093/bjps/axp012.

———. 2013. "Against Structuralist Theories of Computational Implementation." *The British Journal for the Philosophy of Science* 64 (4):681–707. https://doi.org/10.1093/bjps/axs017.

Schmidt, Brandy, and A. David Redish. 2013. "Neuroscience: Navigation with a Cognitive Map." *Nature* 497 (7447):42–43. https://doi.org/10.1038/nature12095.

Shagrir, Oron. 2001. "Content, Computation and Externalism." *Mind* 110 (438):369–400.

———. 2006. "Why We View the Brain as a Computer." *Synthese* 153 (3):393–416. https://doi.org/10.1007/s11229-006-9099-8.

———. 2010. "Brains as Analog-Model Computers." *Studies In History and Philosophy of Science Part A* 41 (3). Elsevier Ltd:271–79. https://doi.org/10.1016/j.shpsa.2010.07.007.

Tolman, Edward Chace. 1948. "Cognitive Maps in Rats and Men." *Psychological Review* 55 (4):189–208.

Vermaas, Pieter E., and Wybo Houkes. 2006. "Technical Functions: A Drawbridge between the Intentional and Structural Natures of Technical Artefacts." *Studies in History and Philosophy of Science Part A*, The dual nature of technical artefacts, 37 (1):5–18. https://doi.org/10.1016/j.shpsa.2005.12.002.

ABSTRACT

THE FALSE DICHOTOMY BETWEEN CAUSAL REALIZATION AND SEMANTIC COMPUTATION

In this paper, I show how semantic factors constrain the understanding of the computational phenomena to be explained so that they help build better mechanistic models. In particular, understanding what cognitive systems may refer to is important in building better models of cognitive processes. For that purpose, a recent study of some phenomena in rats that are capable of 'entertaining' future paths (Pfeiffer and Foster 2013) is analyzed. The case shows that the mechanistic account of physical computation may be complemented with semantic considerations, and in many cases, it actually should.

KEYWORDS: physical computation; semantic account of computation; mechanistic account of computation; mechanistic explanation; causal realization

HYBRIS nr 38 (2017)
ISSN: 1689-4286

WITOLD M. HENSEL
UNIVERSITY OF BIALYSTOK

WATERED DOWN ESSENCES AND ELUSIVE SPEECH COMMUNITIES: TWO OBJECTIONS AGAINST PUTNAM'S TWIN EARTH ARGUMENT

Hilary Putnam (1975) famously contended that the extension of many linguistic expressions is underdetermined by speakers' psychological states taken in their narrow sense, or individuated on the assumption that no psychological state presupposes the existence of any object other than the subject of that state (methodological solipsism).[1] Putnam supported this claim by offering a series of Twin Earth thought experiments and appealing to the phenomenon known as the division of linguistic labor.

In this paper, I focus exclusively on the argument from Twin Earth. I claim that it rests on two assumptions and that these assumptions are highly contentious. Given that both assumptions must be accepted for the argument to work, the argument fails. The first assumption has to do with what Putnam called the logic of natural-kind terms, the second with the notion of a speech community. In what follows, I offer a brief description of Putnam's argument and then focus on each assumption in turn.

"Water" on Twin Earth

The story is familiar. Let there be Twin Earth, a planet that is exactly like Earth in every respect but one: the liquid filling the rivers and lakes on Twin Earth, though (almost) indistinguishable from H_2O, has the chemical composition XYZ. According to Putnam, if a spaceship from Earth visited Twin Earth and its crew discovered the difference between the two planets, the message they would send home would read:

[1] I dispense with the adjective "narrow" in the rest of the paper. The term "psychological state" will henceforth denote narrow psychological states.

On Twin Earth, "water" means XYZ.

The same statement would have been true in 1750, when scientists on either planet were unable to distinguish XYZ from H_2O. Therefore, in 1750, the sentence "Water is tasteless" would have been about H_2O, when uttered by Oscar, and about XYZ, when uttered by Twin Oscar, even if Oscar and Twin Oscar were in the same psychological state (their brains had exactly the same microstructure).

It follows, Putnam says, that knowing the meaning of "water" is not only a matter of being in the appropriate psychological state. Knowing the meaning of "water" also involves having the right kind of causal connections to the right sort of stuff in the world – in this case, to samples of H_2O for Oscar and samples of XYZ for Twin Oscar.

Natural kinds, indexicality and the *qua* problem

The Twin Earth thought experiment would not have been as persuasive as it was with any old word used in place of "water". It could not possibly work with the word "bachelor". And, presumably, a story about Twin pencils, with cores made of some mysterious substance rather than graphite, would not have evoked the response "On Twin Earth, 'pencil' means something else than on Earth", even at a time when *all* pencils on Earth had graphite fillings.

According to Putnam, there is a large class of words, which he calls natural-kind terms, that play an important role in explanations. Natural-kind terms include names of substances, physical magnitudes, animals and plants, as opposed to names of artefacts and other socially constructed objects, such as jobs. They are taken to feature in many inductive generalizations and lawlike statements.

Putnam maintains that they also display a special kind of logic. Namely, invoking a natural kind implies an appeal to a shared (and typically hidden) nature that accounts for manifest characteristics of the kind's members. Thus, any ostensive definition of a natural-kind term carries with it a defeasible empirical presupposition that the indicated sample of the term's extension bears a same-kind relation to most of the stuff to which the term has been applied on other occasions.

Putnam's discussion is somewhat confusing, though. An ostensive definition of any general term carries with it a defeasible presupposition of the type mentioned above. If I say "This kind of writing implement is called a pencil" and point to a pen that merely looks like a pencil, my

definition will not be valid, precisely because the pen I have indicated does not bear the appropriate same-kind relation to objects most members of my language community call pencils. Furthermore, one can define practically any non-empty general term *via* ostension: this is a bachelor, that is a bachelor, etc.

What is relevant to the Twin Earth argument is only that some general terms, including natural-kind names, such as "water", have, as Putnam put it:

> an unnoticed indexical component: "water" is stuff that bears a certain similarity relation to the water *around here*. Water at another time or in another place or even in another possible world has to bear the relation same$_L$ to *our* "water" *in order to be water*. (Putnam, 1975, p. 152)

Therefore, what Putnam calls the logic of natural-kind terms should really be called the logic of indexicality. This is because, on Putnam's account, some names of artefacts have an indexical component, whereas some natural-kind terms *do not*. As to the former, recall Putnam's discussion of Rogers Albritton's live pencils example: If we discovered that pencils on Twin Earth were organisms, we would refrain from calling them pencils, unless of course we discovered that pencils on Earth were also alive (Putnam, 1975, pp. 161-162). "A Mercedes" would be a less controversial illustration.

There are two distinct reasons why it is not the case that all natural-kind terms are indexical on Putnam's view. First, some names of natural kinds, such as "sand" and "air", do not seem to presuppose any particular nature shared by all their referents. This is so, even though we standardly explain the referents' manifest qualities by appealing to microstructure. Second, whether or not a word exhibits indexicality depends on language users. Indeed, according to Putnam, some natural-kind words that begin their career as equivalent to clusters of descriptions can subsequently become indexical (and, presumably, the other way around). For all we know, "water" may have initially meant something like "colorless, tasteless, odorless liquid that quenches thirst".

To complicate things further, Putnam believes that many different senses of the word "water" coexist and many of those senses are indexical. This is because "*X* bears the relation same$_L$ to *y* just in case (1) *x* and *y* are both liquids, and (2) *x* and *y* agree in important physical properties. . . . Importance is an interest-relative notion" (Putnam, 1975,

p. 157). Presumably, the fact that structural properties are "normally" considered to be important implies that, in its *core* sense, "water" denotes H_2O.[2] In some senses, samples of H_2O with impurities are water, in others not; in some senses, ice counts as water, in others not, etc. These differences in extension result from interest-relativity.

What is the relationship between the Twin Earth thought experiments and Putnam's account of indexicality of some natural-kind terms? I suggest that the experiments are taken to confirm the theory, though, of course, cannot establish its truth. Our referential judgments are regarded as the theory's explananda, and indexicality is supposed to account for them. This raises two kinds of questions: about the existence of the target phenomena (e.g., do our referential dispositions comport with the theory's predictions?) and about the theory's ability to explain them, if they exist. In this part of the paper, I focus on the latter kind of questions, so I temporarily grant that our responses to Twin Earth scenarios agree with Putnam's.

There are three major difficulties here. The first is that the world may not have the natural-kind structure required by Putnam's account. Secondly, even if the world has the appropriate natural-kind structure, there are an indefinite number of widely differing, mutually exclusive and equally intuitive construals of reference transmission, the choice of which profoundly affects extension. In other words, in light of the first two problems, Putnam's theory may fail to explain the target phenomena. Thirdly, there may be an account of meaning that explains the target phenomena at least as well as Putnam's account, but in an internalist way. In other words, Putnam's account may not be the best explanation of the target phenomena. Although all these worries are equally important, I will restrict attention to the first.

In a recent paper, Sören Häggqvist and Åsa Wikforss (forthcoming) argue persuasively that the Kripke-Putnam account of natural-kind terms relies on microessentialism, a view according to which objects or samples of substances falling under a single natural kind all share a common microstructure that explains their macroscopic properties and is necessary throughout modal space. But microessentialism, they contend, is at odds with our best philosophy of science. Therefore, the Kripke-Putnam thesis, which asserts that the

[2] Putnam is silent on what makes a particular sense *the core sense*.

extension of natural-kind terms is determined by the microessences of stuff present in speakers' environment, is incorrect.

The core of Häggqvist and Wikforss' argument is well-known to anyone familiar with the so-called *qua* problem and contemporary philosophy of science. Typically, there is no single structure underlying the properties of an object or substance; instead, there are numerous structures that do not fit very well with the natural kinds suggested by common sense. This is especially clear in biology. As Häggqvist and Wikforss poignantly remark, Devitt (2008) is the only author in the philosophy of biology who clings to a form of essentialism. Similar conclusions are being reached by philosophers of chemistry about chemical kinds (see, e.g., Needham, 2000, and Hendry, 2005).

Following Häggqvist and Wikforss, let me summarize what science tells us about the nature of water. First of all, the formula H_2O does not capture the structure of water, but rather its chemical composition, or, in this case, molar proportions. This is an important distinction, because structural isomerism implies that different substances may share a single composition. For example, propanol, isopropanol, or methoxyethane are all C_3H_8O (and large molecules of organic compounds have millions of isomers). Generally, then, chemical composition is not a good candidate for a substance essence – we need to dig deeper.[3] The essentialist may respond by invoking molecular structure: surely, she will say, water is composed of H-O-H molecules, isn't it? Well, not quite. As Häggqvist and Wikforss point out, water is not usually molecular. Liquid water is composed of H^+ and OH^- ions as well as H-O-H molecules, all of which are in constant flux, forming polymers of different lengths at rates that vary with temperature and pressure. On this level of description, then, liquid water has an immense number of structures. And Häggqvist and Wikforss have barely scratched the surface. They haven't broached the subject of heavy, semi-heavy, heavy-oxygen or tritiated water. Nor have they mentioned the fascinating complexities of water in its other states, including different varieties of amorphous ice (LDA, HDA, VHDA).

[3] Water happens not to have isomeric structure, but since other substances do, we should probably look for chemical essences at a lower level of organization than that of chemical composition.

If it is so hard to pinpoint anything even remotely resembling *the* microessence of water, then it should come as no surprise that similar difficulties arise, in much greater numbers, when we turn to biological kinds. It is, I think, no exaggeration to say that modern biology is a thoroughly anti-essentialist science. This anti-essentialism is clearly reflected in the philosophical literature devoted to the life sciences, so I am not going to dwell on it here. Instead, let me recall briefly an interesting study by Andrew Shtulman and Laura Schultz (2008), which suggests that naïve essentialist beliefs about biological species seriously impede people's ability to understand the principles of Darwin's theory of evolution through natural selection. Given the impact of Darwin's theory on contemporary biology, it is no wonder that biologically-informed researchers are so vehement in rejecting essentialism.

Häggqvist and Wikforss's criticism would be potentially devastating against a conception of natural kinds that took singularity of common structure to be necessary for natural kindness and indexicality. Putnam's theory is not that sort of theory, however. Here is a quote that confirms this: "But the local water, or whatever, may have two or more hidden structures – or so many that 'hidden structure' becomes irrelevant, and superficial characteristics become the decisive ones" (Putnam, 1975, p. 1961). The same idea appears in a recent defense of externalism by Daniel Korman (2016), who formulates the following "default conditionals" that are supposed to govern the semantics of "water":

> (i) If water turns out to be compositionally uniform, then "water" expresses a concept that applies to all and only samples of that compositional kind with respect to all counterfactual situations.
>
> (ii) If water turns out to have a highly disuniform composition, then "water" expresses a concept that applies to all and only samples of superficially water-like kinds with respect to all counterfactual situations. (Korman, 2016, p. 507)

Korman's default conditionals are useful, because they wear their shortcomings on their sleeve. First, they are glaringly incomplete. We need at least one more default conditional to handle kinds that are neither uniform in composition nor highly disuniform. Second, the term "highly disuniform" is vague.

We can take care of the first problem by following Putnam, who maintained that the extension of terms such as "jade" is determined by a disjunction of two hidden structures (there are two kinds of jade, he claimed). In fact, we have no other option, since given the apparent lack of natural kinds unified by a single microessence, any other move would amount to adopting internalism rather than externalism.

Now, the vagueness of Korman's phrase "highly disuniform" is a different issue. As I see it, externalists have only two options available to them. The easy way out would be to use the standard method of dealing with vagueness: i.e., draw the boundary between highly and non-highly disuniform kinds in an arbitrary manner. Say, at three or seven, to stick only to magic numbers. This would make the dispute between externalists and internalists a partly conventional and partly empirical disagreement, a matter of decision as well as of fact. Although I do not pretend to have insights into the ultimate nature of reality, I would not bet on the world turning out the way externalists expect it to be. So far, increasing scientific progress has been associated with ever more discoveries of new structures (see Taylor, Vickers, 2017, for an overview of the phenomenon of conceptual fragmentation in science).

The other option is to maintain, like Putnam, that the boundary separating highly disuniform kinds from merely disuniform ones is delineated in light of our interests. Generally, then, the picture Putnam is proposing is that the hidden structures that determine the extension of natural-kind terms are always filtered by our interests. Unless we are talking infinities, the number of structures relevant to explaining what we want to explain in light of a set of interests is bound to be smaller than the number of hidden structures listed in a long unsorted disjunction. Moreover, it is arguably an empirical issue (though in a pretty broad sense of the word "empirical") whether a set of structures is relevant in light of a particular set of interests.

While adding interests to the mix is a step in the right direction, I do not think it will save externalism. Indeed, it will only exacerbate the *qua* problem.

As far as the determination of extension of natural-kind terms is concerned, interests enter the equation in at least two places: when the speaker (or group of speakers) chooses which characteristics of a kind of substance or object need explaining, and when the speaker (or group of

speakers) decides which kind of explanations are acceptable. Needless to say, this is a grossly oversimplified picture of what is really going on.

To illustrate: When you observe water, you regard some of its superficial characteristics as more important than others. For example, you may want to know why water is tasteless or how it is that fish and other animals can live in it, but, at the same time, you can remain unmoved by the fact that water solidifies and increases in volume when the temperature drops below zero. Indeed, most superficial characteristics of water will probably escape your attention altogether. The choice, to repeat, depends on your interests, broadly construed. But, once you have selected which properties of water to account for, your interests will also affect what type of explanations you will pursue and accept. For example, you may prefer functional explanations to mechanistic ones (see Lambrozo, Gwynne, 2014); or you may opt for observation rather than experimentation, because it is cheaper. Although choice of explanans and choice of explanandum are often interdependent, it is reasonable to keep them separate here.

Let me use a toy example to flesh this out. Imagine that you are walking through a jungle and encounter an unfamiliar object or substance – a tree, a shrub, an insect, a mammal, or some malodorous slime oozing from a rock. You study it for a while and decide to give it a general name. You say to yourself "I will call this kind of slime 'shlaw'". You put some of the stuff into a bucket, take it to your village and show it to the shaman, who is visibly excited. Suppose that five superficial characteristics of shlaw become important to people from your village and ten more, though remarked upon, have been largely ignored.

The question to address is this: How have your personal interests and the interests of your community constrained the choice of hidden structures relevant to determining the extension of "shlaw"? Answer: It is hard to say, but probably not very much. First, the characteristics of shlaw that are of interest to you or your community will probably be poorly defined (What exactly *is* slime? What did you mean by "malodorous"?) and thereby amenable to a wide range of theoretical interpretations. This means that, more often than not, they will be discovered by future science to be *clusters* of properties rather than properties *per se*. Second, they will be diverse: each characteristic will most likely be explainable in terms of a different set of hidden structures. And, third, there will typically be a large number of explanatory

approaches acceptable in your community, with each theoretical perspective potentially picking out, via specific idealizations, slightly different structures as explanantia.

Generally speaking, it would seem that knowing the interests of a speech community can help us to identify structures relevant to determining the extension of natural-kind terms only if the community in question is scientifically advanced, for only in such communities can we expect the properties to be well defined and the interests to be sufficiently well articulated. But this is an illusion. The real trouble with interests is that they shift over time, even in scientifically advanced communities. Worse still, these changes of interests are completely unpredictable.

With that in mind, let us try to find out what the word "water" in its "core" sense might denote nowadays. Suppose our present interests and technological development, together with the world, succeed in determining a small set of microstructures underlying the superficial characteristics of water. Can we justifiably maintain that having one of these microstructures is constitutive of water? If so, then what are we going to say when our interests shift, our technology changes and, as a result, a different set of microstructures becomes the most plausible candidate for the nature of water? And if not, then how else should the nature of water be determined?

Given the changeability of interests over time, we can decide that the extension of "water", in its core sense, is determined by the world together with: (a) the interests of our ancestors who first used the word "water" indexically, or (b) our contemporary interests, or, indeed, (c) our future interests – say, the interests of the last generation of our speech community.

Option (a) is implausible, because, as I have already observed, our ancestors' interests were probably too poorly articulated to pick out a sufficiently small number of microstructures. Moreover, there is probably no way of discovering who those ancestors were, what interests they had, and how they used the word "water". And, last but not least, it would be impractical for us to adopt a notion of water that did not harmonize with our present interests.

Option (b) has the obvious advantage of harmonizing with our current interests. However, it does not really bring us much closer to solving the *qua* problem than do accounts of reference that make no

appeal to interests. As things now stand, there are simply too many candidate microstructures to choose from, even if we bring current interests to bear on the choice. Another problem with option (b) is that it is almost indistinguishable from descriptivism, as it practically amounts to asserting that water is identical to whatever satisfies our best current theory. And, just like descriptivism, it also fails to stabilize reference over time: assuming (b), the extension of "water" *has* likely changed since 1750.

Just like (a), option (c) blatantly ignores our present interests. And do we really want the extension of our natural-kind terms to get fixed by something in the future? Moreover, as far as I can see, option (c) can help us solve the *qua* problem only if we adopt a curious form of convergent realism. The convergent realism I have in mind asserts that science will eventually reduce rather than expand the set of microessences plausibly associated with the word "water" (and other natural-kind terms). This, as I remarked earlier, runs counter to the inductive record. Therefore, choosing option (c) would, in most probability, only exacerbate the *qua* problem.

Unfortunately, options (a-c) do not exhaust the possibilities. Not by a long shot. There are indefinitely many accounts we may explore, and many of them would be more appealing than options (a-c) discussed above. But, while I like churning out complex speculative theories as much as the next guy, I will spare myself and the reader the tedium of considering a host of increasingly nuanced accounts of reference. Instead, I will jump right ahead to the conclusions.

Note that options (a-c) are all unsatisfactory, because each tethers the extension of "water" to an arbitrary point in time and thereby imposes unwarranted constraints on acceptable microessences. Option (a) is overly conservative: if our interests are incompatible with the interests of our ancestors, the extension of "water" will probably differ from what we currently take it to be (it is also utterly insensitive to the progress of science). Option (b) is biased in favor of the present and blind to future scientific, technological and social developments. Option (c) anchors the extension of natural-kind terms at the random moment when our speech community will cease to exist. Readers who enjoy apocalyptic books and movies can immediately see the fault in that: what if our civilization collapses and its few survivors, though still speaking English, die out after living for three generations in Dark Age conditions?

All this implies that a plausible account of extension fixing for natural-kind terms must probably involve expanding the set of candidate microstructures associated with options (a), (b) or (c) rather than reducing it.

Let me illustrate this by considering an improvement on option (a). Suppose, for the sake of argument, that our ancestors who first introduced the word "water" as a hidden indexical had interests I_1, whereas, at present, we have interests I_2. Assume also that the set of microstructures constitutive of water as determined by I_1 and the set of microstructures constitutive of water as determined by I_2 have no common element. This means that the extension of "water" as determined by (a) is out of step with current use.

We can remedy this by positing that the extension of "water", though fixed in the past, includes the endpoints of all metaphysically possible trajectories of knowledge development as jointly determined by the natural-kind structure of the world and all possible combinations of human interests. Although much more plausible than option (a), this account yields an indeterminately large number of microessences. However, because the account's plausibility depends on the supposition that it cannot exclude any reasonably acceptable microessences, any credible account of extension fixing for natural-kind terms must satisfy the same desideratum.

To summarize: Given what we know about science, the number of microstructures that can explain the superficial properties of objects or stuff falling under a single natural kind is probably too large to determine the extension of any natural-kind term. It is so, even if we specify the same-kind relation by appeal to interests.

The notion of a speech community

Microessentialism is not the only controversial presupposition of Putnam's Twin Earth thought experiments. A second, though frequently unnoticed, assumption that is involved has to do with the notion of a speech community. This sociolinguistic aspect of Putnam's reasoning was first brought out by Eddy Zemach (1976).

Zemach observed that Putnam's externalist formulation of the imagined report sent from the spaceship back to Earth relies crucially on how speech communities are individuated. If we are liberal and accept

the English speaking inhabitants of Twin Earth as members of our speech community, then the message should read:

We have discovered that there are two kinds of water: H_2O and XYZ.

Since, in Putnam's story, there is exactly as much XYZ as there is H_2O, "water" should presumably refer to a disjunction of H_2O and XYZ. Putnam's description of the situation makes sense only if Twin Oscar does not belong to the same speech community as Oscar.

But what possible reason could we have for excluding Twin Earthians from our speech community, given that *ex hypothesi* the only thing distinguishing them from us is that they happen to inhabit a slightly different environment? Can Putnam mark the distinction between speech communities without begging the question against internalists and excluding Australians, South Africans, or the English?

Zemach is skeptical. Attempts to define language in terms of a speech community are frequently circular, because, more often than not, a speech community is itself characterized as a group of people who speak the same language (see Wardhaugh, 2006). Zemach's worry, then, is that Twin Earthians belong to a different speech community than Earthians, because they speak a different language, and we know that they speak a different language because the word "water" applies to XYZ in Twin English and to H_2O in English.

Zemach's worry is justified. Putnam does not offer any reasons why we should respond to the imagined discovery of XYZ on Twin Earth by saying "The word 'water' on Twin Earth means XYZ" rather than by saying "There are two kinds of water". And the differences between American English and Australian English are both more numerous and more linguistically significant than the alleged difference between English and Twin English. In fact, as Zemach suggests, the idiolects of Oscar and Twin Oscar are probably more similar to one another than the idiolects of Hilary Putnam and any other speaker of American English.

Putnam, however, is not committed to admitting Australians, South Africans, or the English into his speech community. He merely needs to specify a non-question-begging, intuitive method of *excluding* Twin Earthians. Such a method seems available.

It is no profound insight that people belong to speech communities by virtue of communicating with each other using language, among other things. We can exploit this observation to formulate a necessary condition for membership in a speech community:

if a person belongs to a speech community, she must have communicated via language with another member of that speech community. Consequently, two completely isolated groups of people cannot form a single speech community.

The proposed necessary condition does not appeal to the notion of extension or to a particular notion of language and so it is not open to the charge of question-begging. Another advantage is that it may well be intuitive. If it is, then we should expect our spontaneous judgments about the extension of natural-kind terms to vary according to the extent of posited verbal interactions between speakers. As an exercise, consider the following two variations on Putnam's original story:

> (1) In a galaxy far, far away, there is a planet that is almost exactly like Earth. It is inhabited by people that look like exact atom-for-atom replicas of us, but the distance between Earth and the galaxy far, far away is so great that it precludes any causal interaction of the sort necessary for copying. The only difference between Earth and its twin, call it Twin Earth, is that the liquid that fills the rivers, lakes, and seas on Twin Earth, though phenomenologically indistinguishable from H_2O, has the molecular composition expressed by the chemical formula XYZ. Assuming the story is true, does the word "water" (a) refer to H_2O in English and to XYZ in Twin English, or (b) are there two kinds of water, i.e. "water" means H_2O or XYZ?
>
> (2) In a nearby galaxy a long time ago, there was a planet that was almost exactly like Earth. As a result of a cosmic coincidence, it was even inhabited by people that looked like exact atom-for-atom replicas of us. The only difference between Earth and its twin, call it Twin Earth, was that the liquid that filled the rivers, lakes, and seas on Twin Earth, though phenomenologically indistinguishable from H_2O, had the molecular composition expressed by the chemical formula XYZ. About a thousand years ago, a spaceship from Earth visited Twin Earth, and, at the same time, a spaceship from Twin Earth visited Earth. Having discovered each other's *Doppelgängers*, Earthians and Twin Earthians began travelling back and forth, talking on the radio, writing letters, etc. Only recently, and to their great astonishment,

> have they discovered the difference between the two planets. Assuming the story is true, does the word "water" (a) refer to H_2O in English and to XYZ in Twin English, or (b) are there simply two kinds of water, i.e. "water" means H_2O or XYZ?

Your answers will count as evidence for the intuitiveness of Putnam's externalism if your confidence in (a) is noticeably higher in scenario (1) than in scenario (2). I confess that my own responses agree with Putnam's.

But even if most readers' answers to scenarios (1) and (2) agreed with Putnam's and mine, this would merely establish that we seem to share some beliefs about speech communities. It would not secure the stronger conclusion that the beliefs in question are intuitive in any interesting sense of the word. Since intuitive beliefs are standardly construed as strongly influenced by our biological makeup, it is useful to think of intuitions as deeply engrained domain-specific assumptions about the world. What makes these assumptions interesting is the strong causal connection between their etiology and our evolutionary history. This is why intuition are taken to be practically universal across cultures.

It is doubtful, however, that there is a folk theory that relies on an intuitive notion of a speech community. We cannot claim the same kind of familiarity with the inner workings of speech communities as we can with the behavior of water, animals, and individual people. We hardly make – and we practically never *have to* make – any inductive generalizations involving speech communities in everyday life. Consequently, it is very difficult to identify any relevant features shared by speech communities besides the necessary condition proposed above. Lastly, unlike the words "water" and "animal", the phrase "speech community" has a distinctly theoretical ring to it. In sum, Noam Chomsky's (2000, p. 148) famous criticism of Putnam's account applies directly to the notion of a speech community:

> We can have no intuitions about the question, because the terms *extension*, *reference*, *true of*, *denote*, and others related to them are technical innovations, which mean exactly what their inventors tell us they mean: it would make as little sense to explore our intuitions about tensors and undecidability, in the technical sense.

This suspicion is borne out by how the term "speech community" has been used in empirical linguistics (see Patrick, 2002, and Wardhaugh, 2006, pp. 119-132). The first modern definitions of "speech community" appealed primarily to uniformity of linguistic behavior. For example, Bloomfield (1933) explicitly assumes that "within certain communities successive utterances are alike or partly alike", and then adds, by way of a definition, that "any such community is a speech community" (Bloomfield, 1933, pp. 153-154). Bloomfield's emphasis on uniformity is later echoed in Chomsky's "ideal speaker-listener, in a completely homogenous speech community (Chomsky, 1965, p. 3; see also Chambers, 1980). Interestingly, according to Patrick (2003), Bloomfield explains both external boundaries and internal variation in terms of speaker interactions: "a speech-community is a group of people who interact by means of speech" (Bloomfield, 1933, p. 42) and "differences of speech within a community are due to differences in density of communication" (Bloomfield, 1933, p. 46). In other words, on Bloomfield's view, Twin Oscar would belong to the same speech community as Oscar and the similarities between their utterances would remain inexplicable. Needless to say, Chomsky and his followers would whole-heartedly agree.

The advent of sociolinguistics, ushered in by William Labov's presentation at the 1962 annual meeting of the Linguistic Society of America (see Chambers, 2002, p. 5), did not change much when it came to the theorists' lack of reliance on speaker interaction. Although some sociolinguistic approaches appear to be consistent with Putnam's externalist account, because they either retain the interaction condition (Gumperz, 1968)[4] or characterize speech communities in terms of geographic location, most accounts in the field appeal to criteria that have nothing to do with density of communication. Perhaps the most influential such criterion invokes shared norms of utterance production and evaluation. For example, Labov writes (1972, p. 120-121):

> The speech community is not defined by any marked agreement in the use of language elements, so much as by participation in a set of shared norms. These norms may be observed in overt types of evaluative behavior, and by the uniformity of abstract patterns of variation which are invariant in respect to particular levels of usage.

[4] See the next paragraph.

A striking feature of the many notions of a speech community explored in sociolinguistics is that none of them unequivocally classifies Oscar and Twin Oscar as members of different speech communities. Even notions closest to Putnam's proposal fail to do so. For example, Gumperz's definition, according to which a speech community is "any human aggregate *characterized by regular and frequent interaction* by means of a shared body of verbal signs and set off from similar aggregates by significant differences in language usage" (Gumperz, 1968, p. 381, emphasis added) either treats Oscar and Twin Oscar as members of the same speech community or, at best, leaves the matter unsettled, because Americans do not interact with Twin Americans, and yet the utterances made by Oscar and those made by Twin Oscar are linguistically indistinguishable.

In fact, even the most noncommittal definition of a speech community I know of, which stipulates that a speech community is merely "some kind of a social group whose speech characteristics are of interest and can be described in a coherent manner" (Wardhaugh, 2006, p. 119), would arguably be of no use to Putnam, because Oscar and Twin Oscar share all verbal dispositions, being linguistically indistinguishable from each other.

Of course, it is fairly easy to modify Putnam's story so that Twin English becomes a distinct language from English. Just introduce a sufficient number of differences in pronunciation and perhaps syntax. But there are two problems with this move. First, it is now unclear whether Oscar and Twin Oscar are in the same psychological state, because, at the very least, their brains are no longer identical. Second, the string of words represented as "Oscar would like a glass of water" should not count as a single sentence, but rather as two: one in English and one in Twin English. Indeed, if Putnam insists that Oscar and Twin Oscar belong to two different speech communities, the most accurate report sent back to Earth should read:

People on Twin Earth use a word that sounds like the English word "water" and applies to a substance that looks and behaves exactly like water, but its chemical composition is XYZ.

Yet the fact that two different words may denote two different substances is old news. It is hardly a profound insight that the word "water" means water, whereas the word "fire" means fire.

Externalists can shift gears, however, and insist that the Twin Earth story yields a desired conclusion concerning words, not meanings. The conclusion would be that knowledge of words does not supervene on psychological state, because Oscar and Twin Oscar, though psychologically indistinguishable, know different words. Alas, this argument faces a similar problem to the previous one. Namely, in order for the externalist conclusion to follow, Oscar and Twin Oscar must be in different brain states if one is to belong to a different speech community than the other. And although the externalist can reply that people who are in two different brain states may well be in the same psychological state, she will need an additional argument for the psychological irrelevance of neurological properties underlying linguistic differences.

Externalists can also point out that sociolinguistics is not a mature field, and it has not yet produced a good enough notion of a speech community. It is therefore possible that a mature sociolinguistic theory will recognize Twin Oscar as belonging to a different speech community than Oscar. But this misses the point. For, regardless of how sociolinguistics will develop, the important thing is that the notion of a speech community is the kind of concept that is shaped by the investigator's interests. As Patrick puts it (2002, p. 593):

> we ought not to assume SpComs [speech communities – W.M.H.] exist as predefined entities waiting to be researched or identify them with folk notions, but see them as objects constituted anew by the researcher's gaze and the questions we ask.

Ultimately, then, both externalism and internalism are viable positions in so far as their choice is informed by the researcher's interests. If, however, we choose not to ignore current scientific practice when assessing philosophical positions, then externalism appears to be the less plausible alternative.

REFERENCES

Chambers, J. K. (1980). Linguistic Variation and Chomsky's "Homogenous Speech Community". In M. Kinloch, A. B. House (eds.), *Papers from the Fourth Annual Meeting of the Atlantic Provinces Linguistic Association* (1-32). Fredericton: University of New Brunswick.

Chambers, J. K. (2002). Studying Language Variation: An Informal Epistemology. In J. K. Chambers, P. Trudgill, N. Schilling-Estes (eds.), *The Handbook of Language Variation and Change* (3-13). Oxford: Blackwell (1st edition).

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Devitt, M. (2008). Resurrecting Biological Essentialism. *Philosophy of Science*, 75, 344-382.

Genone, J., Lambrozo, T. (2012). Concept Possession, Experimental Semantics, and Hybrid Theories of Reference. *Philosophical Psychology*, 25 (5), 714-741.

Gumperz, J. J. (1968). The Speech Community. In D. L. Sills (ed.), *International Encyclopedia of Social Sciences* (381-386). New York: Macmillan.

Häggqvist, S., Wikforss, Å. (forthcoming). Natural Kinds and Natural Kind Terms: Myth and Reality. *The British Journal for the Philosophy of Science*.

Hendry, R. F. (2005). Lavoisier and Mendeleev on the elements. *Foundations of Chemistry*, *7*, 31-48.

Korman, D. Z. (2016). What Externalists Should Say About Dry Earth? *The Journal of Philosophy*, 103(10), 503-520.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lambrozo, T., Gwynne, N. Z. (2014). Explanation and Inference: Mechanistic and Functional Explanations Guide Property Generalization. *Frontiers in Human Neuroscience*, 8, Article 7.

Needham, P. (2000). What is Water? *Analysis*, 60(265), 13-21.

Patrick, P. L. (2003). The Speech Community. In J. K. Chambers, P. Trudgill, N. Schilling-Estes (eds.), *The Handbook of Language Variation and Change* (573-597). Oxford: Blackwell (1st edition).

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

Shtulman, A., Schulz, L. (2008). The Relation Between Essentialist Beliefs and Evolutionary Reasoning. *Cognitive Science*, 32, 1049-1062.

Taylor, H., Vickers, P. (2017). Conceptual Fragmentation and the Rise of Eliminativism. *European Journal for the Philosophy of Science*, 7(1), 17-40.

Wardhaugh, R. (2006). *An Introduction to Sociolinguistics*. Oxford: Blackwell (5th edition).

Zemach, E. (1976). Putnam's Theory on the Reference of Substance Terms. *The Journal of Philosophy*, 73 (5), 116-127.

ABSTRACT

WATERED DOWN ESSENCES AND ELUSIVE SPEECH COMMUNITIES: TWO OBJECTIONS AGAINST PUTNAM'S TWIN EARTH ARGUMENT

The paper presents two objections against Putnam's Twin Earth argument, which was intended to secure semantic externalism. I first claim that Putnam's reasoning rests on two assumptions and then try to show why these assumptions are contentious. The first objection is that, given what we know about science, it is unlikely that there are any natural-kind terms whose extension is codetermined by a small set of microstructures required by Putnam's indexical account of extension determination. The second objection is that there may not be a plausible concept of a speech community whose adoption would classify Oscar and Twin Oscar as members of different speech communities and, at the same time, render Oscar and Twin Oscar as being in the same psychological state. I contend that Putnam's argument fails because both objections are justified.

KEYWORDS: externalism; Twin Earth argument; natural-kind terms; qua problem; interest relativity; speech community

HYBRIS nr 38 (2017)
ISSN: 1689-4286

KATARZYNA KOBOS

WHAT DOES THE SENSORY APPARATUS DO WHEN THERE IS NOTHING TO PERCEIVE? THE SALIENCE OF SENSORY ABSENCE

An adequate and exhaustive explanation of the mechanism behind perception should cover both typical and anachronous instances of sensory acts under standard circumstances as well as in conditions deviating from the established norm. Subject to the discussion herein should be situations when a perceptual act occurs in the absence of sensory stimuli. There is no unequivocal resolution as to what happens in consciousness processing the lack of perceivable sensory qualities. Let us tackle the problem by sketching two different models of perceptual response to the absence of sensory signals.

The first model: Stimuli fuel the sensory apparatus. Whatever cannot be sensed, can be deducted. Inference of the absence of sensory stimuli

The first model stipulates that perceptual activity phases out with the shortage of stimuli. Since there is nothing to sense, sensation simply does not occur. If nothing excites the receptors, there is nothing to trigger sensations. This is, thus, a model that affords a minimal contribution of higher cognitive features to perception. Along these lines, the role of the senses in recording external environment goes no further than to relay the excitation by some form of external energy, be it visible radiation, an acoustic wave, etc. Pursuant to the model, there is a straightforward answer to the question of the function of the perceptual apparatus when there is nothing to sense and physical stimuli fail to achieve the absolute threshold required for their detection by the sensory receptors. The solution is: perception does nothing. Perception is restricted to the

registration of positive sensual qualities. However, since consciousness still stands or rather its stream briskly flows onwards even in the idle mode of the sensory functions, it remains to be brought out where the awareness of lacking sensory qualities resides if perception is not the one to grasp gaps in the incoming cascade of signals from ambient environment. This point undermines the model. The above line of reasoning proves the reaction-dependent model of sensory experience to be discriminatory in the sense that successful perception occurs only under steady external input. The model's explanatory power dissolves in the absence of positive sensory qualities. In order to restore credibility to the model, the lack of sensory qualities would have to be attributed directly to the subject. If there is nothing to see, this unperceivable absence cannot be referred to anything in the outside world and it is inherited by the subject herself. But can we defend the contention that the lack of sensory signals may be construed to fall within the domain of sensory experience? Under the conjecture that vision ends when there is no more to see another sensory faculty would have to take over, with proprioception being the only candidate left on the battleground. Ludicrous as it sounds, all missing components of experience would have to be considered as forged by proprioception. This would further lead to the conclusion that darkness or silence would be detected by inner sense organs and would belong to a different category than positive visual or acoustic sensations. Such an assumption cannot be upheld. Under the restrictions of the model, the absence of a specific type of sensory signals invokes higher cognitive faculties. You cannot see darkness or hear silence but there is no obstacle to inferring that darkness or silence is around. This model postulates an interventionist role of higher cognitive faculties with respect to sensory experience as they step into action only when lower level faculties face resistance. The model seemingly ascribes higher autonomy to the non-sensual component of experience but it invites confusion as to why abstract thinking and other forms of discursive operations can only be at stand-by under standard perceptual circumstances. The adoption of the model requires acknowledging a discrete, intermittent mode of operation of higher cognitive faculties in the perception channel as it is their task to ensure the continuation of perceptual activity upon a gap in the influx of sensory signals. The shortage of external stimuli may only be conceived of in abstract terms.

This brings us to the crux of the conceptual unease spawned by the reaction-only model. Granted that missing stimuli, regardless of their sensory channel, equal non-perception, it is unclear on what basis the subject could discriminate between the respective commonly identified types of perceptual absence. Now the lack of acoustic signals seems very conspicuous and has been dubbed silence, whereas the scarcity of visual stimuli goes by the name of darkness, the dearth of flavor is nowhere near the unavailability of odor. Zero gravity feels differently than an interruption in the flow of tactile sensations. There also exist states that do not readily yield to categorization, such as painlessness that seems devoid of specific constituent features. Is there a model that would do justice to the variations of missing sensory qualities? Conscious interpretation of the state of deactivation of sensory receptors must be revisited in further models.

Let me make one explanatory remark before it is contested that receptors are fueled by some form of energetically pumped signals from the outside. No matter what the actual neural activation pattern is, what I shall refer to as a receptor for the sake of this text should be scaled down to whatever organelle, substance, or structure that directly undergoes stimulation and absorbs the energy. It might not always be a cell. Even though photoreceptor cells, such as rods and cones, may undergo inhibition when hyperpolarized by light and release neurotransmitters when not hit by radiation, it does not challenge the fact that either there is external input or there is none. Confusion may only arise if we fail to clearly delineate what we mean by a receptor.

A counterargument against full dependence of perception on external stimuli stems from the consideration of the conceptual distinction between the inability to take in a specific type of stimuli, such as deafness, and the recording of a state conspicuous for the lack of signals from the environment, such as silence. Now the difference disappears if no line is drawn between sensory perception, activation of a sensory organ, awareness of the existence of a specific type of sensations, and the actual presence of a definite type of sensory stimuli in the environment. Some researchers point to the fact that we are often at a loss as to whether we have just forfeited hearing or silence has fallen until we are exposed to further acoustic stimuli. (Sorensen 2008, s. 2). A common denominator of these states is surely the lack of signal from the auditory organ. But what is salient for the conscious subject is not so

much the fact that nothing is to be heard, but why it is so. The source of disturbance in the reception of acoustic stimuli is what matters in terms of survival and effective action. The organism should discriminate between its own sensory failure and a null level of the intensity of a specific type of stimuli. The explanatory and discriminatory futility of the model is manifested by the unavailability of a clear distinction between deafness and the sensation of silence.

This calls for a recasting of the posed question concerning the function of the perception of sensory qualities in terms of external world representations. The outside world is not out there for the sake of being subject to perception. Perception serves for the organism to stay a part of the world and survive by navigating within it. Not only what stimuli impinge on the sensory organs is of significance. Of equal or far surpassing importance is to determine why exactly their influx has stopped. The recognition of perceptual absence falls short of the task of staying alive unless the context of the omission is embraced. We can, by all means, speak of a continuous sensory deprivation of a person suffering from an irrevocable loss of specific sensations in the wake of damage to either sensory organ, neural path leading to the brain, or the respective cerebellar area itself. It may be doubted, however, that it is sound to claim that such a person is constantly exposed to silence. She is deprived of acoustic sensations on equal terms with a person with intact hearing but staying in a completely soundproof room. However, it is a transient experience for the latter person who expects a deluge of acoustic signals once she leaves the room. It shall not be an overstatement to claim that such a person hears silence, whereas the deaf person, in fact, fails to hear. The state of missing sensory qualities does not only represent physical reality but is also crucial for a successful interpretation of the course of events.

The advocates of representationalism in the treatment of the status of conscious experience, such as Michael Tye, could repeal the objection concerning the lack of a representationalist component of silence or darkness by contending that such a component is indeed present and corresponds to the idleness of respective sensory organs. In effect, a sensation of silence or darkness would carry information pertinent to the state of sensory organs and not to the absence of phenomena in outside reality. (Tye 2003, p. 166) If we were to be consistent, however, we would have to assume on this view that we

exclusively perceive the states of sensory receptors both upon their excitation and upon deactivation, never reaching out deeper into the outer world. This would compromise the very concept of sensory perception. We perceive in order to navigate outside environment. If sensory perception is to retain its rationale, it should represent external reality and not the states of stimulation of sensory organs.

It seems, however, that a representation generated without the contribution of an external factor and designed exactly to stand for its absence may only be construed broadly as spanning a wide fragment of reality. A singular simple sensation appears to be ineligible for representing missing impingement of the environment on a specific type of sensory receptors. Only a non-atomistic, complex, manifold representation may render the deficiency of a chosen quality. The corollary is that there is no representation of silence, or darkness, or any other lacking sensation at the level of the sensory apparatus, and its identification requires the comparison of two representations, one sound-laden, the other - soundless. On this view, silence as a separate individual aspect may not be represented by the receptors. Although the initial state of missing stimulation is identical in the case of both sensory deprivation and the absence of sensory stimuli, brain-level representations may vary.

We reach the heart of the problem at this point. It seems fully justified to claim that silence may be heard. It must have the status of a sensation in order to fulfill its dedicated function. How to reconcile the status of a representation of the external environment with a full-blown sensation in case of negative sensory qualities? Is it possible that sensation is not confined to the domain of sensory organs? To what extent is the brain privy to the birth of sensory experience and how much does it conspire in the way the world seems to us? In what sense is the brain an accomplice in the formation of sensory experience and not only its recipient?

How far can we take the contention that the absence of sensory stimuli does not translate into the lack of experienced sensations but constitutes a separate type of perception? How are we to interpret the fact that sensory void is accompanied by perceivable neural correlates, negative sensory qualities? Silence does not stand as a state of completely damped acoustic stimuli. On the contrary, it designates a specific sensation, divergent from deafness in a constitutive way and not

only genetically, by dint of its cause. The analysis of the phenomenon of darkness approximates exactly such an account in the sense that darkness may be counted among other sensations. It was Aristotle who already identified darkness with black, i.e., one of three achromatic colors alongside with white and grey. Modern explanation of the sensation of black does not depart from the original ancient finding, and it is thought to be the response of the visual apparatus to the lack of electromagnetic waves in the visible bandwidth (Hurvich 1981, p. 61). The bottom line is whether a theoretical backdrop may be developed wherein sensory experience would be traceable to higher cognitive faculties.

The first model is not to be pronounced doomed altogether. What should be questioned is its pertinence to conscious perception. However, the model holds true for sensory receptors, adequately capturing their function, which consists in the detection and differentiation of stimuli. Without the perceptual apparatus in place, there is no point in tracking the emergence of sensory experience. But episodes when external stimulation discontinues do not involve sensory cell excitation. It is not the receptors that record the lack of sensory qualities. Negative sensory states are detected based on the absence of the output from sensory organs but they are identified only at higher level cerebellar structures. It remains to be determined at exactly what level. The analogy between senses and sensors that indeed record both positive and negative states of a chosen aspect of the environment - collapses. These devices operate in a binary way. This is their intended use. In the case of complex organisms such as ourselves, we may safely assume that it is the central nervous system that serves as the recording and storage medium of the unavailability of specific sensory qualities and not the subordinate sensory organs. A substantiation is called for at this point that the discussion concerns the lack of sensory stimuli and not the exclusion of a sensory aspect from the scope of attention. Sometimes, the disappearance of a sensory quality, what can and cannot be seen or heard, is not due to the actual state of the external environment, nor the dysfunction of receptors, nerves, or the respective information processing areas in the brain, but other global brain mechanisms impacting conscious thought. A varying share of awareness may be apportioned to distinct inputs, depending on their salience. Even a distinct sound may fall beyond the scope of consciousness if it becomes

too monotonous a signal for the brain to bother. Maybe, we only grasp the transition from an environment bathed in acoustic waves to silence, with the outcome being that we are only aware of changes. (see (Maruszewski 2001, pp. 62–67)) The absence of a conscious sensation does not necessarily mean sensory stimuli have failed to occur. This higher-level selectivity of perception has come to be called attention. It may well be just another side of perception. (Lupyan, Clark 2015, p. 282) What I address in this paper are the episodes when the lack of stimuli is in the focus of attention.

Feedback perception model: Reversal of the order of sensory perception. Perception as a continuous process of forecasting the environmental inputs by higher cognitive faculties assisted by the perceptual apparatus upon a mismatch of the top-down prognosis and the actual state of the matter

To address the second model we must revisit the heretofore introduced category of representation. Allow me to add that this applies to a representation at the level of neural tissue, regardless of the exact mechanism behind the coding (internal neuron composition, global cerebral reach of select neural connections, specificity of neural synapses, oscillations of electromagnetic waves at a specific frequency inside the brain, etc.) While this point is beyond the scope of this study, let me note in passing that the very claim and subsequent evidence that a fraction of the world, namely, the brain, carries a host of finely structured hierarchical neural representations of the world beyond and within the body (inner bodily affairs) poises us to discard the now redundant stipulation of a supplementary category of mental representations. If neural representations are forged through interaction with the world itself in a succession of generations, the mind need not internalize the world again.

In the discussion herein it is assumed that the senses consistently respond to specific types of external stimuli. This condition trivially assures the effectiveness in representing the environment. A stimulus of type X always invokes a reaction of sense a, while stimulus of type Y always triggers the response of sensory function b. It is pointless to ask about accuracy in representing the environment as access to physical reality is granted via the senses and higher cognitive faculties, so there is no external point of reference. As long as sensory experience ensures

survival and effective action, it may be considered that experience is indeed accurate – in a trivial sense. The more faithfully the structures and functions of the organism encode the external considerations, the higher the odds for survival and evolutionary success. The accuracy or correspondence of experience to the ambient environment, the resolution of this mapping, is the work of evolution and factors shaping neural machinery through the succession of the forms of life. It may be worthwhile to recast the assessment of experience in terms of precision. The question is not whether the representation of external reality in experience is true, but if it holds relevance. The precision of experience would consist in such an attunement of senses to the environment that they would serve not as signals for detection and interpretation but a clear call for action. Imagine a sensory landscape in need of further thorough processing, providing a complex map to be read according to a detailed legend. Such a far removed sensory interface with the world is possible to navigate but hardly effectively. It lacks immediacy. A much better design would allow us to read signals from the world as signposts and warning alerts that are readily actionable.

While the accuracy of representation is granted, precision may be perfected also throughout individual development, by means of practice. What is more, precision only makes sense in terms of an organism's interaction with the environment. Passive staring at an object fails to serve a specific purpose and hence it does not fulfill the criterion of fitness. It is far from being a representative example of an organism's activity. What organism is solely engaged in perception? Perception usually accompanies intricate movements, keeping balance under challenging conditions, compensation of undesirable shift of the center of gravity through the flexion or extension of extremities, complex manipulation of objects. Its natural settings are those of motion and relocation of the organism within its environment. Organism herself fine-tunes her senses – through rearrangement of bodily location, posture and gestures, eyeline shift, etc. Perception is not available in a read-only mode, it is an editable interface with the external world.

Such a conception renders it impossible for experience to be born here and now only to be reborn a moment later, in every single instant of conscious life. Should it be that way, the subject would be constantly engaged in perception alone. If experience seems seamless and appears to carve the world at its joints, it must have been finely sculpted by the

cognitive machinery of the brain in the course of evolution. Its close alignment with external reality comes as no surprise. It has emerged in conjunction with reality. To represent the external environment, experience does not have to reproduce it each and every time, as it is reality's own extension, and reality has guided its formation and contributed to its development.

A question springs to mind: since the brain builds the image of reality so precisely, why does the organism ensure the constant activity of the senses? The image of reality (Clark 2015, p. 3) cascades with top-down traffic of forecasts and requires only some adjustments of the senses in case of a deviation from the prognosed flow of sensations. Alas, I have not authored this model. One of its advocates is Andy Clark, and the model is dubbed embodied predictionism. On this account, the brain is a prognostic device projecting the expected train of events amid the stream of sensations. Higher cognitive faculties bear the brunt of the burden and their projections are only so much as validated by the signals from the sensory receptors (Clark 2015, p. 5). Hence, most of the receipt of sensory signals occurs below the threshold of awareness. It is only upon a prognostic error that the senses come to the surface of consciousness and the direction of cognitive processing is reversed from receptors to higher brain areas. The tightly-knit frame of expected experience readily accommodates a direct sensation, the operation of external stimuli themselves. Errors in the prognosis of unfolding experience provide a window onto external reality. Under normal circumstances, which also explains why it seems to be the standard, the forecast conjectured by higher cognitive faculties prevails, undisturbed by perceptual discrepancies.

I have sketched the theory of predictionism. It is now time embodied predictionism is put to the test in the context of the lack of sensory stimuli.

How does this theory account for the perception of missing sensory stimuli? Let's reframe the question in more detail. It goes without saying that an exclusive comprehension of individual negative sensory qualities gives rise to a cognitive discord. Imagine an attempt to narrow down all sensory channels to just the receipt of the absence of, let's say, sounds. How could you possibly experience solely silence or contemplate deafness with the attenuation of all other sensations? That said, the reconciliation of a single missing type of sensory quality with

the overall representation of a vast momentary perceptual landscape invites few objections. The sensory repertoire seems richer and is not limited to singular qualities. A collective perceptual act of a full sensual stage, albeit deprived of acoustic or visual sensations, occurs commonly and allows to avoid the paradox of directly seizing something that isn't there. In a further step, also forecasting such rich multi-ingredient sensory vistas, devoid of individual types of sensory qualities, satisfies all criteria of credibility. It is clear that a conscious subject shall expect darkness at night and shall not be surprised by the fact that there is nothing to excite her rod and cone cells. Except perhaps for a lone photon.

What seems counterintuitive and requires a thorough theoretical analysis is the defense of the embodied account of perceiving the lack of sensory qualities. Here, embodied predictionism encounters a true challenge and opportunity for deploying the depth of its explanatory potential. Negative sensory qualities may be represented, they may also be predicted. But will they succumb to the account of embodiment? The idea behind embodiment rests on the deliverance from the need to build a succession of finely detailed models of the world in individual instants of experience for the sake of putting in place a stable precise representation of the external environment in the brain as well as interaction with reality. A simpler, less challenging explanation of the conformity and convergence of the actions undertaken by conscious creatures and the actual state of the matter indicates not as much an incredible ability to seamlessly represent outside reality as the involvement of reality itself in the workings of the sensory apparatus. Neither at the receptor level nor at the level of higher brain structures is there room for a real-time, ongoing strict mapping of the world at a suitable resolution. However, there is room for a matrix of probable responses and the selective activation of this matrix causing specific sensations. The senses and higher brain areas need not elaborate baroquesque world representations in each fleeting moment. Rather, an evolution-licensed cognitive template is set against the ambient circumstances and only interferences must be accounted for, if there are any. Thus construed, perception is an interplay with the world and not an act of collecting inputs to be encoded so that other parts of the brain may unpack them and process them further. The organism takes advantage of a sophisticated matrix corresponding to the outside

environment that forecasts the unfolding of experience. It is none other than the brain along with its integral neural correlates of phenomena (Clark 2015, p. 4). Since the bulk of sensory details comes from our own prognostic-representation machine, should there emerge an inconsistency between the forecast and the actual stimulus, the subject is poised to stand face-to-face with a single sensory stimulus, something unthought-of in traditional representationalist work where all perceptual interactions were considered to be heavily mediated (Feldman&Friston 2010,p. 2). An interference with the prognosis makes a direct appearance in consciousness, rather than being passaged via the entire convoluted interpretation and decryption pipeline of perceptual brain areas, which would impair immediacy. To ensure the readiness of capturing a random element gone against the strain of the top-down forecast, the influx of stimuli must be met with an equally complex correlate so that no other detail burdens the processing channels but this one outlier. Another prerequisite is the active manipulation of this correlate against and within outer environment – through motion, adjustment of bearings, etc. This is what embodied cognition is all about. The organism may incessantly tweak its vantage point, by tilting the head, moving around in space, squinting eyes. This makes sensory cognition a dynamic and multidimensional act. Perception does not occur statically, we do not come to learn about the world from aloft, from the position of a remote observer, but from the inside. The senses do not have to simulate the external world as their calibration with the world occurs in real time through actual motion, adjustment of posture and bodily position in space against external objects.

How are we then to interact with something that isn't there? The subject must deploy a model of states of sensory deprivation. But such a model indeed exists and is inscribed in the brain structures. There is thus no need for the cognitive functions to form a representation of negative sensory qualities on each individual occasion. Under the assumption that the subject is equipped with rich resources representing the world of which it is an extension, amassed through long-term exposure of its ancestors and itself, the paradoxicality of the experience of sensory absence is dissolved. It remains a fact that if no acoustic waves reach the environment, there are no visible electromagnetic waves to occur, there are no tactile stimuli, sensory cells stay idle. The respective types of sensory deprivation may not be encoded at the level of sensory receptors

as it is precisely them that are not involved in the least. But they are not the carrier of consciousness, and the brain that takes advantage of their machinery by no means succumbs to idleness. After all, it manages to determine with precision what sensory cells fail to evince activity. The mechanism behind this relies on the same principles that account for external excitations triggering respective responses at the level of conscious experience. If specific neurons fall silent, isn't it a sufficiently clear signal to release an appropriate sensory response? Why shouldn't gaps in the flow of a specific type of sensory qualities be interpreted as direct experience? If the elaborate correlate of external reality covers also negative sensory qualities, why shouldn't the subject detect the absence of specific stimuli in the environment under suitable circumstances? Directly. The lack of sensory qualities does hold the status of experience. It is the subject herself that integrates such qualities into the image of reality even though they are absent in the environment. In this respect, the subject's repertoire of sensory qualities exceeds the one available among external stimuli. We may indeed see darkness.

The account of embodied predictionism strives to demonstrate its empirical viability. Much as a raft of evidence has been submitted in favor of the soundness of predictionism (see, e.g. Friston 2011), embodiment is yet to be demonstrated more thoroughly in empirical settings (although a body of relevant studies exists - see Beer 2000, p. 97; Glaescheri et al., 2010, p. 585). In pursuit of research methodology and an empirical trial of the theory, let us resort to a thought experiment.

Let us consider a unique situation of a person with a condition fit to be called „negative synesthesia". Such a person manifests sensitivity to all sensory stimuli that invoke a conscious response in a human being adequately equipped for her species. It, therefore, comes as no surprise that this person can also properly identify the lack of stimuli of any type. The specificity of perception in such a person consists in that the sensation of silence is accompanied by a visual experience of darkness, the deficiency of tactile stimuli leads to the feeling of odourlessness, etc. We could point out all the feasible ways the senses may interfere with one another, according to the principles of combinatorics, but let me leave it at the examples provided. Now let us now turn to the thought experiment. Let us assume that a negative synesthete enters a dark room reverberating with the chords of the piece *Kind of Blue* by Miles Davis. The hapless gal recognizes in the sequence of sounds one of her favorite

musical compositions but acoustic bliss is spoiled by the ambient darkness that she is susceptible to experiencing as silence. The cognitive state of a negative synesthete may be preliminarily described as that of dissonance. Contrary to visual and acoustic illusions and other standard cases of paradoxical perceptual acts, such as the detection of motion in a static drawing, or the attribution of various dimensions to objects with identical size due to the impact of contextual cues on the interpretation of sensory data, or hearing ever rising tones in a repeated sequence of music, negative synesthesia does not result from some kind of a conflict in higher level processing of sensory information and applies to baseline sensory qualities. It should be noted in passing that impairment of one of the senses usually causes the amplification of the function in others. The very phenomenon of negative synesthesia is thus scarcely probable in nature. It may, however, aid the discussion herein. How can you hear silence in the accompaniment of acoustic sensations? In the light of a standard model of bottom-up sensation, whereupon signals ascend from the receptors to brain areas without significant feedback from the brain, the receipt of sound stimuli excludes a sensation of silence. This is one and the same sensory channel. An analogous regularity pertains to all other sensory modalities. Please bear in mind that under this model two allegedly conflicting experiences occurring in the same sensory channel, i.e., the sensation of silence and audible tones of the melody *Kind of Blue*, belong to two separate orders. On the one hand, the representation of silence forms due to atypical stimulation of perceptual brain areas. On the other, the latter experience emerges as acoustic waves impact the sensory cells of the hearing apparatus. The paradoxicality of the experience of silence in the accompaniment of music dissolves if contradictory sensory components are assigned to disparate categories. The disintegration of sensory consistency is at its highest in case of negative synesthesia, with the model of bottom-up perception failing to rule out such a possibility.

Whereas on the grounds of predictionism, neural correlates also feature the representations of negative sensory qualities. It is thus possible for the brain, busy with forecasting the course of experience, to collate them with positive sensations. In principle, the case of negative synesthesia makes sense under the model of perception woven by the prognostic brain machinery, faced with the actual inputs of sensory stimuli. It remains an open issue what combinatorial algorithm could

serve for the cognitive apparatus to merge the tissue of experience. What would be the reaction of a brain identifying darkness with silence and simultaneously exposed to acoustic stimuli in a darkened room? Would a negative synesthete hear a hushed melody? Perhaps a negative synesthete would enjoy a full-bodied auditory sensation only in a well-lit room with light-colored walls and equipment, and she would encounter hearing difficulties in dimmed lighting? Before we delve into far-fetched speculations, let us evoke the temporarily disregarded aspect of embodiment. Predictionism in its own right fails to undermine the feasibility of negative synesthesia. Will embodiment prove more restrictive? Negative synesthesia implies that each occurrence of missing sensory stimuli would be accompanied by total sensory deprivation. However, with the emergence of the first tones of *Kind of Blue* the brain should align the feeling of severance from the signals from the outside world and reinterpret it as irrelevant to the auditory channel. Since hearing receptors detect acoustic waves, there is no room for the sensation of silence in the organism's cognitive economy. It is the external environment itself that serves as the referee in all unequivocal perceptual scenarios. The non-neurotypical brain structure of a synesthete may invoke the sensation of silence in response to the lack of visual stimuli, but further interaction with the environment and the intake of reverberating sounds should dismiss the unsubstantiated experience of silence. Whenever anything deviates from the prognosis, the subject witnesses the making of perception. The absence of observations of behaviors indicative of the hypothetical phenomenon of negative synesthesia comes as no surprise and serves as evidence in favor of embodied predictionism.

REFERENCES

Clark, A. (2015). *Embodied Prediction*, in: Metzinger, T. & Windt, J.M. (ed.), "Open MIND".

Lupyan, G. & Clark, A. (2015). Words and the World: Predictive Coding and the Language-Perception-Cognition Interface, *Current Directions in Psychological Science*, vol 24, no. 4, pp. 279-284.

Hurvich, L. M. (1981). *Color Vision.* Sunderland, Massachusetts: Sinauer

Keeley, B. (1999). Fixing content and function in neurobiological systems: the neuroethology of electroreception. *Biology and Philosophy* 14: 395–430.

Maruszewski, T. (2001). *Psychologia poznawcza*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.

Sorensen, R. (2008) *Hearing Silence: The perception and introspection of absences*, Forthcoming in *Sounds and Perception: New Philosophical Essays*, ed. Matthew Nudds and Casey O'Callaghan, New York: Oxford University Press.

Sorensen, Roy (1999). Blanks: Signs of Omission. *American Philosophical Quarterly* 36/4: 309–321.

Sorensen, Roy (2007). *Seeing Dark Things.* New York: Oxford University Press.

Tye, M. (2003). *Consciousness and Persons*. Cambridge, Massachusetts: The MIT Press.

ABSTRACT

WHAT DOES THE SENSORY APPARATUS DO WHEN THERE IS NOTHING TO PERCEIVE? THE SALIENCE OF SENSORY ABSENCE

This study aims to bring out the explanatory potential of embodied predictionism versus passive feed-forward model of sensory stimulation in the pursuit of a parsimonious naturalist account of sensation as a salient feature and an end point of conscious experience. Theoretical approaches towards sensory experience are tested against specific scenarios of the absence of observable or palpable qualities including but not limited to the thought-experimental phenomenon of negative synesthesia at the conclusion of the argument. . Predictionism is first explored in its own right only to be found insufficient to do justice to the actual mechanism behind full-blown immediate perception. A case is made for the soundness of predictionism reconciled with the doctrine of embodiment.

KEYWORDS: embodied predictionism; representation; embodiment; sensory deprivation; absence of sensory stimuli; sensory perception

HYBRIS nr 38 (2017)
ISSN: 1689-4286

MAREK POKROPSKI
UNIVERSITY OF WARSAW

MENTAL CONCEPTS: THEORETICAL, OBSERVATIONAL OR DISPOSITIONAL APPROACH?

1. Problems of other minds - introduction

It is not necessary to articulate the problem of other minds in detail since it is one of the classical problems in philosophy of mind (for an overview, see Avramides 2001). The problem has three formulations: ontological, epistemological, and, the most recent one, conceptual. In a nutshell, the ontological problem of other minds concerns the existence of other minds: do other minds exist? Epistemological problem raises the question of the possibility of knowledge of other minds' content: how can I know what others think or feel? The key question of the epistemological problem is whether mental states are private in nature and thus unobservable or, on the contrary, they can be perceived directly. According to the former, I can directly access only my mental states but not others'. The others' mental states cannot be observed directly, therefore they must be cognized in some indirect way. The latter claim gives possibility to ground knowing other minds on perceptual basis.

Since René Descartes, several different solutions of the epistemological problem of other minds have been proposed, including such influential positions as inference from analogy (e.g. J.S. Mill, A.J. Ayer). Development of cognitive sciences in the last few decades and research in the field of social and developmental psychology renewed the debate (the

so called Theory of Mind debate) introducing a body of empirical research, which resulted in new positions. However, in this article I will not discuss all positions in the debate. There are simply too many of them to consider in this short paper. Moreover, recently in the debate there has been an attempt to reconcile indirect approaches (such as theory-theory or Simulation Theory) with the direct perception account and thus propose a hybrid theory (e.g. Fiebich & Coltheart 2015, Carruthers 2014, Stich & Nicols 2003). In general, hybrid theories acknowledge that we have more than one cognitive strategy of "mindreading", for instance, perceptual and inferential, which we use depending on different factors. For example, the default strategy would be perceptual, and inference would be used second in case of insufficient perceptual information (Carruthers 2014).

The third formulation of the problem of other minds is the conceptual problem, the origin of which can be found in Wittgenstein's late philosophy (Wittgenstein 1968, Avramides 2001). This problem concerns the possibility of acquiring mental concepts, such as pain or sadness, that are universal, i.e., mental concepts which could be equally ascribed to myself and others. If we grasp the nature of pain on the basis of our "inner" experience, then how can we ascribe this concept of pain to others? To put it differently, how can mental concepts, which we understand on the basis of our experiences, be used both in the first as well as in the third person cases? A negative solution to the problem is to acknowledge that we have two different mental dictionaries, one first-personal and the other third-personal. This idea is not only counterintuitive but also generates the problem of similarity criteria between concepts from different dictionaries. The challenge, then, is to develop a plausible positive account of mental concepts—one that accommodates the application of concepts in both the first and third person cases.

Not all proposals in the contemporary theory of mind debate address the conceptual problem. Thus, in the article I will focus only on these approaches which either consider the origin and nature of mental concepts explicitly, such as theory-theory (TT), or implicitly, like Direct Perception (DP), dispositional or phenomenological approach of Merleau-Ponty.

2. Theory-theory and Direct Perception

According to theory-theory (TT) we can know what others think or feel on the basis of inference (e.g. Baron-Cohen 1995, Carruthers 1996, Stich 1983). We infer mental state of the other when we perceive his or her behavior by employing a theory (folk psychology) about other people's mental lives and behavior (Stich & Nicols 2003). Folk psychology can be understood here in two ways: (1) as a set of skills of mindreading, that is, skills of attributing mental states and predicting the others' behavior, or (2) as a collection of platitudes or a set of generalizations regarding the mental life of others and causal relations between mental states and external stimuli; e.g., if someone receives a painful stimulus, then he/she feels pain, which results in screaming, crying, etc. (behaving in a specific "painful" way). One can argue that it is impossible to give a list of all folk-psychological platitudes. Although that is surely true, it is not necessary. Folk-psychology platitudes are rather putative, tacit, commonsensical knowledge, which is used implicitly in the process of mindreading. One can also raise the question: how do we acquire these platitudes? Some theory theorists (e.g. Carruthers 1996) claim that at least some of them are innate. Others argue that we gain them during development and through the acquisition of cultural practices.

The direct perception (DP) account claims that at least in some situations we can directly perceive others' mental states such as intentions and emotions (e.g. Cassam 2007, Gallagher 2008, Smith 2010a). The question of how perceptual knowledge of other minds is possible remains. Different versions of the DP account provide different answers. For example, Joel Smith (2010a) argues for a perceptual account using the Husserlian concept of perceptual co-presentation and a functionalist approach to mental properties. Seeing others' mental states would be similar to perceiving three-dimensional objects: only the front side is sensually "present", but the back side is perceptually "co-present". Smith admits, however, that it is plausible that in different situations we have different strategies of gaining knowledge about others, including perceptual and inferential strategies. Quassim Cassam argues differently for the perceptual model. He claims "that one can sometimes know what

others are thinking or feeling by visual means" (2007, p. 170). This solution is supported by Dretske's (1969) theory of epistemic perception which is non-inferential.

Now, how these approaches address the conceptual problem of other minds? According to theory-theory, mental concepts such as pain, sadness, or belief are theoretical terms, which at some point were introduced into our folk psychology. We use these theoretical entities to explain and predict behavior of others as well as our own. It may seem that TT omits the conceptual problem by postulating a common dictionary of mental terms and a common folk psychology. But how do we introduce and define these theoretical mental terms? Theory theorists usually refer here to functional definition. The idea of functional definition of theoretical mental terms was proposed by David Lewis (Lewis 1970, 1972; Stich 1983).

> "Call these *theoretical terms* (*T-terms* for short) because they are introduced by a theory. Call the rest of the terms in the story *O-terms*. They are all the *other* terms except the T-terms; they are all the *old*, *original* terms we understood before the theory was proposed. We could call them pre-theoretical terms." (Lewis 1972, p. 88-89)

To illustrate his idea, Lewis tells a detective story (1972). In the story, the detective investigates the death of Mr. Body. The detective observes the crime scene and notices various phenomena such as the victim's body, blood on the wall, a broken window, etc. Then he proposes an explanation of the mystery, introducing the story of three individuals called X, Y, and Z who conspired to kill Mr. Body. The detective describes what role X, Y, and Z played in the conspiracy and the act of killing. When the detective is introducing his story, he does not know the real names and nature of X, Y, and Z, they are theoretical terms defined by their functional role. Their real names can be discovered in further investigation, if the theoretical hypothesis is true.

By analogy, mental concepts are theoretical terms introduced in order to explain human behavior. We use them to explain and predict others' behavior as well as our own. Mental concepts as theoretical terms

are defined functionally, that is, by their functional role they play in the cognitive system. They are not observational terms, but observational terms (such as stimuli or bodily responses) can be used in their definitions. Lewis agrees that folk psychology was never introduced in a specific moment in the history of science, which makes it difficult to differentiate pre-theoretical terms from theoretical ones. Thus, he acknowledges that folk psychology is a myth, however, as he argues, it is a good myth because it gives us plausible explanation of social cognition.

Besides the mythical origin of theory-theory, there are other problems with the functional definition of mental concepts, such as "narrow causal individuation" (Stich 1983, pp. 22-23). In short, causal individuation means that mental states are determined only by their causal interactions. theory-theory holds the narrow version of causal individuation, which means that causal links which determine mental states, are only those between mental states and other mental states, between mental states and stimuli, and between mental states and bodily responses or behavior. This means that functional definitions of mental terms are narrow and explanations produced by theory-theory cannot include links that go far beyond the organism, for example, past events or sociocultural facts. This obviously constrains explanatory power of TT, especially in highly contextual cases of human behavior.

The next objection raised by Stich (1983) concerns causal links between mental states and behavior. Theory-theory claims that particular mental states, say, the experience of a headache, typically cause particular behavior, say, taking painkillers. However, this is only a statistical law dependent on one's age, knowledge, social status, and, say, susceptibility to the pharmacological industry. Thus, "typically causes" is highly variable and dependent to various factors, which TT cannot address due to narrow constrains.

Finally, it seems that TT omits the conceptual problem by postulating the same set of mental terms introduced in folk psychology and used to explain others' as well as our own behavior. It is claimed that the grounds of self- and other-ascription are basically the same, namely, inference to the best explanation. It is not clear, however, if in both cases we deal with the same *explanandum*. In the case of other-ascription, data

are clearly behavioral, we explain what we actually see from a third person perspective, whereas in self-ascription cases, it is highly plausible that we deal with a sort of inner first-personal experience or introspective data. If so, then we use the same set of mental terms, defined using third-person terms (e.g. observational) to explain different phenomena, both first-personal and third-personal. Another solution is that the grounds of self- and other-attribution do not have to be the same. For instance, according to Carruthers, it is plausible that other-attribution is based on "inference to the best explanation of (behavioral) data", whereas "self-knowledge should be thought of analogous to the theory-laden perception of theoretical entities in science" (Carruthers 1996, p. 26). Accordingly, self-attribution is a kind of non-inferential (at least at a personal level) recognition of one's mental state, which is characterized in mental (theoretical) terms.

Now let's consider how the Direct Perception account addresses these issues. DP states that we grasp others' mental states on a perceptual basis, i.e., in direct observation of someone's behavior. The cognitive process behind it is considered to be non-inferential but requiring a conceptual content for mental attribution. However, mental concepts are not theoretical terms, but they come from perception, and thus can be understood as either observational terms or ones that are reducible to them. In strong interpretation of DP, mental states are identical with behavioral states. This, however, generates the conceptual problem. How can we know that our mental concept of pain denotes identically the same behavioral state of the other? One way to answer this question is to reject introspection or any other kind of "inner" access and acknowledge that self- and other-ascriptions are grounded on the same basis, namely external observation (e.g. Cassam 2007). In some limited cases, it is plausible that we ascribe mental states by observing ourselves. But even if that is the case, the observational access from the first-person perspective and third-person perspective are radically different. According to Joel Smith (2010b), the direct perception account does not solve the conceptual problem of other minds. Moreover, it generates an analogous conceptual problem of other bodies, i.e., we end up with two separate sets of concepts

of behavioral/mental states, one from the first-person perspective the other from the third-person perspective.

A weaker version of DP holds that relation between "inner" mental states and "outer" behavior is more complex. For example, Overgaard and Krueger propose a different reading of direct perception which redefines the relation between bodily expressions and mental states (Overgaard & Krueger 2012). They defend Direct Perception account referring to phenomenologists such as Max Scheler and Maurice Merleau-Ponty and argue that bodily behavior is "constitutive" of mental states, which means that "certain bodily actions make up proper parts of some mental phenomena" (2012, p. 257). According to that, "we see others' emotions by seeing proper parts of their emotions" (p. 255), which are embodied and observable. To use Overgaard's and Krueger's example, the tip of an iceberg is in this sense a proper part of iceberg and it might be said that seeing the tip of an iceberg on the horizon is to notice that there is an iceberg. It is not clear, however, what "constitutive" means here and how it is different from just "being a part of". Tip of an iceberg is a visible part of the iceberg, similarly to the front side of a chair I see in front of me. If so, then maybe, following Smith (2010a), it is better to consider this relation in terms of co-presence and apperception instead of "constitution". Furthermore, even if we agree that we can grasp mental states via "proper parts", we do it either by external observation or by a sort of "inner" experience (e.g. proprioceptive experience of facial expressions, which are proper parts of an emotion). Thus, such interpretation of direct perception does not help to solve the conceptual problem. Still, the mental terminology is divided between the first-personal and the third-personal. In order to give plausible account of conceptual problem, this dichotomy has to be overcome.

3. Dispositional and phenomenological account

Choosing between theoretical and observational terms is not a satisfying solution for the conceptual problem of other minds. Both theory-theory and direct perception do not solve the problem but, moreover, they generate more problems. Is there a third option? There is at least one

interesting candidate, in favor of which I would like to argue. This account conceives mental concepts as dispositional terms.

Dispositional account is usually linked with behaviorism, for example with Gilbert Ryle (1949/2009) and thus is a sister of direct perception. According to Ryle, mental concepts have dispositional nature i.e. they refer to subject's dispositional properties. When we call someone intelligent or melancholic we express that he or she has tendency to behave in a particular way when specific conditions are realized. For example, we would call someone intelligent if he or she, when asked, answered questions concerning general knowledge. Importantly, dispositions concern not only what we actually observe but, first and foremost, what we would see when specific conditions were realized. Thanks to dispositional concepts we are able to foresee what will happen and explain what happened. Accordingly, mental concepts are dispositional terms which we use to predict and explain others' behavior.

Development of this approach was recently proposed by Eric Schwitzgebel (2013), who introduces dispositional account of attitudes. Schwitzgebel argues that:

> "to have an attitude is, (…) to have a dispositional profile that matches, to an appropriate degree and in appropriate respects, a stereotype for that attitude (…) To have an attitude (…) is mainly a matter of being apt to interact with the world in patterns that ordinary people would regard as characteristic of having that attitude." (Schwitzgebel 2013, p. 75)

To generalize this claim: to have an attitude, belief, or to have an emotion or feeling, such as pain, is to behave accordingly with a stereotype for that belief, emotion or feeling, or, as Schwitzgebel puts it, to "live a certain way" (2013, p. 76).

The key notion of this approach is stereotype. According to Schwitzgabel "a stereotype for a property X is a cluster of other properties that would be regarded as characteristic of something that possesses property X" (2013, p. 81). Not all properties are equally important for a stereotype, some are more, other are less. Thus, stereotype can be conceived as a space of properties from which some are more central,

other are peripheral. This approach specifies dispositional concepts as a piece of commonsensical knowledge which comes from folk psychology. For example, if someone believes that it is going to rain, he or she will wear a raincoat or take an umbrella. If someone has pain in his/her knee, he or she will limp, walk slowly, take painkillers etc. It seems, however, that dispositional terms cannot be reduced to observable data, because they concern all possible behavior matching the stereotype. Moreover, some behavior is highly contextual and depends on environmental and cultural conditions. This advantage lead at the same time to difficulties e.g. the acquiring problem (how do we know which properties constitute a stereotype?) and the selection problem (which properties form the stereotype cluster are central?). Simple answer states that we know all of this from folk psychology and present context. However, as I showed above, folk psychology has difficulties with narrow causal individuation, that is, in putting mental terms in socio-cultural context and long-time dependencies. Indeed, in some cases cultural background and personal history as well as bodily knowledge of skills can have strong influence on explaining behavior of others and ourselves. If so, then maybe it is worth trying to replace folk psychology with another approach.

3.1 Phenomenological account

Phenomenological account of other minds, especially the existential phenomenology of Merleau-Ponty, is often read as a version of direct perception (Gallagher 2008, Overgaard & Krueger 2012). Here I would like to argue for a slightly different reading, namely that Merleau-Ponty's explanation of intersubjective cognition is similar, to some extent, to the dispositional account.

First of all, Merleau-Ponty argues that the ontological and epistemological problem of other minds are results of false dualistic ontology, which existential phenomenology is going to overcome. Mental states are not "inner" and private in the sense that they are not accessible for others. They are private only in the sense that we have first-personal access to them. Others, however, can have a third-personal access to my mental states and vice versa. This third-personal access, however, is not mediated by a theory. For Merleau-Ponty, understanding others' mental

states is not a theoretical enterprise but a bodily practice. Thus Merleau-Ponty, even if he claims something similar to the dispositional approach, he would oppose explaining cognition of others using theoretical terms of folk psychology. This does not mean, however, that social cognition does not have conceptual content. Mental concepts shape our understanding of others but have experiential basis. In *Phenomenology of perception* Merleau-Ponty writes:

> I perceive the other as a piece of behaviour, for example, I perceive the grief or the anger of the other in his conduct, in his face or his hands, without recourse to any 'inner' experience of suffering or anger, and because grief and anger are variations of belonging to the world, undivided between the body and consciousness, and equally applicable to the other's conduct, visible in his phenomenal body, as in my own conduct as it is presented to me. (1945/2005, pp. 414-415)

Merleau-Ponty's solution to the epistemological problem goes like this: we do perceive mental states, such as anger or grief, in other's behavior but they can be grasped only as instantiations of structures of existence or "belonging to the world". These structures of existence are anonymous, yet experienced as living body, they are neither first-personal (self-consciousness) nor third-personal (material body). To understand one's intention, grief, or sadness is to apprehend a certain variation of existential structure, which we all share. These structure has many dimensions including: emotional attunement, intentional action, language. For Merleau-Ponty all of them are embodied and intertwined. A change in one dimension, say, a mood change, affects other aspects, say, temporality of action, or linguistic or gestural expressions.

Let's consider an example. We see someone holding his or her knee and limping towards a bench. The perceived movement, facial gestures etc. express not only the intention and objective of action (to find a place to sit), but also its affective mode. The hurting knee shapes the subject's sensorimotor pattern and thus reconfigures situatedness in the environment. We perceive someone's limping movement as expression of pain and intention – looking for relief. However, what we apprehend is not isolated "inner" feeling of pain, but a holistic bodily disposition. On this

basis, we expect a certain set of behaviors and thus we can predict what observed person is up to. The situation is similar in the case of emotional states. When we see someone is afraid, say, of a spider on the wall, we grasp not only a particular object of fear, but the disposition to act in a specific way, say, to scream, move in the opposite direction, ask for help etc.

An important difference between Merleau-Ponty and dispositional approach is that he emphasizes interactive, practical, and embodied nature of social cognition. As he writes in *Phenomenology of Perception*:

> No sooner has my gaze fallen upon a living body in process of acting than the objects surrounding it immediately take on a fresh layer of significance: they are no longer simply what I myself could make of them, they are what this other pattern of behaviour is about to make of them. [...] now, it is precisely my body which perceives the body of another, and discovers in that other body a miraculous prolongation of my own intentions, a familiar way of dealing with the world. (1945/2005, pp. 411-412)

Merleau-Ponty argues that we not only understand what others do and could do in the environment, but also, and maybe most importantly, how we can interact as agents. The other is not a theoretical entity which I have to construct with theoretical terms but an embodied agent in whom I see a "familiar way of dealing with the world".

Accordingly: i) I understand the other's behavior because I share the same existential structures (such as attitudes, emotions, sensorimotor capacities) which shape bodily experience; ii) understanding the other is based on the primal recognition that the other is also an embodied subject; iii) apprehension what the other feels, thinks, does etc., is an apprehension of his/her existential disposition or, to put it differently, an actual way of living; iv) apprehension of other's disposition is immediately connected with my own dispositions, beliefs, and possible actions. I understand the other's behavior through myself and vice versa. I learn about myself thanks to others.

Now, having this background, how can we answer the conceptual problem? From Merleau-Ponty's perspective, mental concepts are not

mental in the sense of being first-personal, inner and private, but they concern certain modes of existence, or, to put it differently, shared dispositions of being in the world. Being in fear, is neither a peculiar "inner feeling" given in first-personal experience, nor a belief "in the head". Fear is a mode of emotional attunement with the world, and being in fear shifts different aspects of experience: it shapes bodily movements, gestures, thoughts, as well as practical engagement with surroundings. In short, being in fear changes our relation to the world on multiple levels.

Merleau-Ponty's account can be read as an extension of dispositional account, however, disposition is understood here in a wide existential sense. It concerns our bodily and affective situatedness in the environment. Mental concepts would be dispositional terms understood as a multimodal (e.g. visual, motoric) representations of behavior. There is a threat, however, of misinterpreting such representations in internalist way. For example, according to Vittorio Gallese, we can read Merleau-Ponty's phenomenology of intersubjectivity in terms of embodied simulation (Gallese 2005). Gallese argues that we use neuronal representations of behavior in an internal simulation process, which results in mental ascription to others. There are, however, serious doubts whether this interpretation of phenomenological account is valid (Zahavi 2012). Another reading, the so-called interaction theory, argues that social understanding is rooted in bodily practice of social interaction, which is understood as a dynamic and co-regulated process between autonomous embodied agents (e.g. Froese & Gallagher 2012). Accordingly, mental concepts would be minimal models of interaction which are deployed and specified in context of particular social interaction.

In sum, to be afraid, greedy, or hungry means to act, think, and feel accordingly with a specific behavioral profile (stereotype). Our understanding of such profile and applying relevant concept in everyday situations depends highly on the context, our previous experiences, as well as on sensorimotor capacities. This means that despite the fact that we share mental concepts as representations of social interaction and thus can understand each other, our experiences are not identical – to put it simply, your pain will never be my pain, although I understand what it is like to be

a subject of painful experience and I know possible profiles of behavior related with such experience.

4. Conclusion

The problem of other minds emerged from the Cartesian framework, where minds were considered as inner, isolated, and self-evident entities. The problem with mental concepts has the same origin. If we accept the view that mental states are "inner" and unobservable, like theory-theory, then we have to acknowledge that mental terms are theoretical constructs, although useful in explaining behavior. If we accept the possibility that we can, at least in some cases, see what others feel and think, then mental concepts have perceptual basis. Dispositional account, at least in the standard version above, argues for the dispositional nature of mental states, but it inherits some problems and constraints of theory-theory and folk psychology. Phenomenological reading of dispositional account argues for experiential and embodied basis of mental concepts used in social cognition, which primarily is social interaction. This approach not only gives justice to the complexity of social cognition and experience of others but also explains dispositions as situated in an environment and embodied.

REFERENCES

Avramides, A. (2001) *Other Minds*. Routledge.

Baron-Cohen, S. (1995) *Mindblindness: An Essay on Autism and Theory of Mind.* Cambridge, MA: MIT Press.

Carruthers, P. (1996) *Simulation and self-knowledge: a defence of theory-theory*. In P. Carruthers and P.K. Smith (eds.) *Theories of Theories of Mind*. Cambridge, Cambridge University Press.

Cassam, Q. (2007) *The Possibility of Knowledge*. Oxford University Press.

Dretske, F. (1969) *Seeing and Knowing*. London: Routledge & Kegan Paul.

Fiebich, A. & Coltheart, M. (2015) *Various Ways to Understand Other Minds. Towards a Pluralistic Approach to the Explanation of Social Understanding*. "Mind and Language" 30(3), pp. 235-258.

Froese T., Gallagher S. (2012) *Getting interaction theory (IT) together*. "Interaction Studies" 13 (3). pp. 436-468.

Gallagher, S. (2008) *Direct perception in the intersubjective context*, "Consciousness and Cognition", vol. 17.

Gallese, V. (2005). *Embodied simulation: From neurons to phenomenal experience*. "Phenomenology and the cognitive sciences" 4. pp. 23-48.

Lewis, D. (1972) *Psychophysical and Theoretical Identifications*. "Australasian Journal of Philosophy" vol. 50, pp. 249-58.

Lewis, D. (1970) *How to Define Theoretical Terms*. "The Journal of Philosophy", Vol. 67, No. 13, pp. 427-446.

Merleau-Ponty, M. (2005/1945) *Phenomenology of Perception*. tr. C. Smith. Routledge.

Overgaard, S. Krueger, J. (2012) *Seeing subjectivity: Defending a perceptual account of other minds*. In S. Miguens and G. Preyer (eds.), *Consciousness and Subjectivity*, pp. 239-262. Heusenstamm: Ontos Verlag.

Ryle, G. (1949/2009) *The Concept of Mind*. Routledge.

Smith, J. (2010a) *Seeing Other People*. "Philosophy and Phenomenological Research". Vol. LXXXI No. 3.

Smith, J. (2010b) *The Conceptual Problem of Other Bodies*. Proceedings of the Aristotelian Society, Vol. cx, Part 2.

Stich, S. (1983). *From Folk Psychology to Cognitive Science*. MIT Press.

Stich, S., Nicols, S. (2003) *Folk Psychology*. In Stich. S. & Warfield, T. A. (eds.). *The Blackwell Guide to Philosophy of Mind*, Oxford: Basil Blackwell. pp. 235-255.

Schwitzgebel, E. (2013) *A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box*. In *New Essays on Belief. Constitution, Content and Structure*. (ed.) Nottelmann, N. Palgrave Macmillan UK. pp. 75-99.

Wittgenstein L. (1968). *Philosophical Investigations*. Blackwell.

Zahavi, D. (2012) *Empathy and mirroring: Husserl and Gallese*. In: R. Breeur and U. Melle (eds.), *Life, Subjectivity & Art: Essays in Honor of Rudolf Bernet*. Springer.

ABSTRACT

MENTAL CONCEPTS: THEORETICAL, OBSERVATIONAL OR DISPOSITIONAL APPROACH?

In the article I discuss the conceptual problem of other minds and different approaches to mental concepts. Firstly, I introduce the conceptual problem and argue that solutions proposed by theory-theory and direct perception approach are inadequate. I claim that mental concepts are neither theoretical terms nor observational terms. Then, I consider third option which states that mental concepts are dispositional terms, i.e. they concern particular patterns (stereotypes) of behavior. Finally, I argue that dispositional approach is to some extent coherent with phenomenological account and that phenomenological concept of embodiment can improve this position.

KEYWORDS: other minds; concepts; theory-theory; direct perception; dispositions; phenomenology

HYBRIS nr 38 (2017)
ISSN: 1689-4286

PRZEMYSŁAW R. NOWAKOWSKI
INSTITUTE OF PHILOSOPHY AND SOCIOLOGY
POLISH ACADEMY OF SCIENCES

EMBODIED COGNITION: LOOKING INWARD

> *Being embodied is being able to take risk, that is, being open and exposed to the unknown.*
> (Depraz 2005, p.173)

Introduction

Lawrence Shapiro, in his book *Embodied Cognition* (2010), distinguishes three types of relations between standard research on cognition and the embodied cognitive science: conceptualization, constitution, and replacement. In the first case, our conceptualization of the world is body-based. In the second case, the body and even some artifacts constitute cognition. In the third case, cognition should be explained in terms of embodied, ecological dynamical systems. As Shapiro (2010) and other commentators point out (e.g. Wilson, Golonka, 2013), only the last one is incompatible with research in standard cognitive science, and at the same time, it is the only really interesting option. Seven years after the publication of this book, it is quite clear that the replacement hypothesis is far from being successful (see: Goldinger et al., 2016). And, as noticed, even if it succeeds, it often fails to explain phenomena that are traditionally called cognitive (Aizawa, 2014; 2015a; 2015b).

This paper unfolds the view which integrates a computational and embodied approach to cognition (see also: Rupert 2016; Miłkowski 2016). However, I assume it here without argument. I argued for an integration of embodied and computational view on cognition somewhere else (Nowakowski, 2017). Still, many authors point out the role of action or interaction, body experience, or artifacts in cognition,

but more detailed works on internal processing are still rare (however, see: Allen, Friston, 2016; Clark, 2013; 2015; Miłkowski, 2016; Rupert, 2016; de Bruin, Michael, 2017). Therefore, I propose some introductory, empirical considerations on internal, cognitive processing in bodily cognitive systems.

In this paper, I start with remarks on internal, cognitive processing. After that, I refer to Alvin Goldman's moderate approach of embodied cognition (highly incompatible with replacement hypothesis). Goldman (2012; 2014) considers the crucial role of body representations (B-codes) in cognition. So, we can propose some remarks not only on internal processing but perhaps also on the role of body representations in this processing.

After the remarks on Goldman's approach, I will sketch my own approach (E-codes approach), based on some conceptual twists. Crucial for embodiment of cognition will be not the role of the body for cognition, but—as I will argue— the role of cognition for the body. This is nothing new, but sadly it is a still too-often neglected view on this matter (but, see: Haselager et al., 2008; Keijzer, 2015). This twist may lead to integration between work on embodied cognition and evolution of the nervous system. After all, embodied cognitive systems are mainly bodily machines, living organisms coping with problems they face in their own surroundings. In the view proposed here, E-codes should be efficient, robust, and body-specific.[1]

Relating E-codes to Goldman's approach, we can say that the central nervous system is undoubtedly an essential part of the system responsible for cognitive processing. However, there is a reason to believe that such a system can extend beyond the boundaries of the brain (see: Nowakowski, 2017; Wilson, 2010). Therefore, we can differentiate:

B-codes: Body related processing;

E-codes: Efficient, robust, and body-specific processing.

Therefore, we can ask:

a. Is it possible that B-codes are E-codes?

b. What conditions must be met for B-codes to be a kind of E-codes?

[1] This issue will be elaborated in more detail later in the text.

I am strongly convinced and I argue that we should bind E-codes (not pure B-codes) with embodied cognition. Therefore, as was mentioned earlier, we should not ask what the body does for cognition, but what cognition does for the body. Hence, I start with some evolutionary considerations on cognition, and then relate this to considerations on the role of the body in shaping cognition.

1. The complexity thesis and the internal processing in the embodied cognition

I begin my remarks with cognition, and the body in which cognition is embodied, then I move to complexity theses and the work of Keijzer with Arnellos (2017, and Keijzer, 2015) on the evolution of cognition (Godfrey-Smith, 1996) with more recent works on the evolution of the nervous system, and I propose a more internalist view on the evolution of cognition. They argue for the important role of both environmental and bodily complexity. After that, I turn to initial remarks about internal processing in embodied cognition.

1.1 Cognition that is embodied

For our purposes, we can reuse a part of the title of Aizawa's paper (2015b): "What is this cognition that is supposed to be embodied?" There is an ongoing debate both outside and inside the research on embodied cognition about what cognition could be. Currently, there is strong criticism that in research on embodiment we deal not with cognition but behavior misdescribed as cognition (Aizawa, 2014; 2015a; 2015b). So what is cognition?

Here I refer to interesting remarks from Buckner's (2015) paper. He writes:

> "[…] cognitive scientists should collect the behaviors that they are interested in explaining as the result of cognition. They should then theorize about a minimal set of capacities that would allow systems to display these behaviors, and see whether agents possessing capacities that allow them to pass one set of behavioral tests also tend to possess the others. If it is plausible that they do, then scientists should attempt to develop a model of the underlying mechanisms that could produce those capacities and explain why they

> would tend to cluster together" (Buckner, 2015, pp.310-311).

Therefore, basic cognition (in our terminology: cognitive processing) is realized by a cognitive mechanism. This mechanism gives the system a set of capacities for the realization of some cognitive behaviors. Behaviors are the effect of employing a cognitive mechanism. In the same paper, Buckner (2015) describes cognition as an ecumenical, homeostatic cluster of properties such as context sensitivity, fast adaptation, grouping/categorization, abstract learning, multi-modality, inhibition, and monotonic integration. As far as this is the cluster, cognition doesn't need to exhibit all mentioned properties in order to be cognition. Certainly, in cases of minimal cognitive processing, it can contain only some of these properties.

I assume that cognition is based on the information processing process of problem-solving. This process should be context sensitive, rapidly adapting to new problems, and related to categorization and inhibition, but it not need be abstract or multimodal. We can also mention that it may not need to be representational. Here I follow some works of Keijzer and his colleagues (Keijzer 2003; van Duijn et al., 2006), assuming that cognitive processing is problem-solving, embedded in sensorimotor coordination and interaction with the environment (Keijzer, 2003). I don't reduce the whole cognition to sensorimotor coordination, but try to show that cognition is something primary related and submerged in this coordination. Some of this coordination requires solving some environmental and body complexity problems. Therefore, minimal cognition is a problem-solving process embedded in sensorimotor coordination. Of course, as cognition becomes more sophisticated, more elements from Buckner's (2015) cluster should be included.

1.2 Embodiment of the cognition

It seems that what we, as theoreticians of the embodiment, should be particularly interested in is the extent to which the body, excluding the central nervous system, is part of the cognitive system (see the definition of embodied cognition in Wilson and Foglia [2011]). This is undoubtedly an important and by no means trivial question. However, in this paper, I focus on internal processing. As argued earlier in embodied

cognition cognitive processing, the base should extend beyond the central nervous system (Nowakowski, 2015). However, the body never independently performs cognitive processes but co-realizes cognition together with the central system (Nowakowski, 2017). Therefore, we deal here with a system characterized by a trade-off between what the peripheral nervous system and non-neuronal body parts do, and what the central system does. I defended the view that in many simpler systems peripheral systems play a greater role in base realizing cognition, whereas in the case of more complex systems (e.g. mammals) the central nervous system plays a greater role in cognition. Some authors (Fuchs 2011; Jacob, 2012; Gallagher et al. 2013), driven by the need of simplicity in cognitive processes, believed that one should conceptualize cognition as depending as much as possible on the peripheral system, whereas periphery should make cognitive processing simpler. Naturally, this will simplify the complexity of central processing. However, sometimes the simplest solution, in general, is to increase dependence on the central processing. There are possible types of embodiment where cognition relies mostly on central processing.

Here, I believe that we can connect this approach with Wilson and Golonka's (2013) idea that "to explain cognition we should focus on a specific task and their sources used during the task." Undoubtedly, among the resources an essential element is the central system; in the case of many animals, it's the central nervous system and we should be able to show what this system really does. Therefore, in this context we can ask: How important is the brain as a resource for the bodily cognitive system or as a central processing machine? Even if an exhaustive answer is not available, we should be able to say what condition central processing should meet to be a part of the bodily cognitive system. I will return to this issue when discussing Goldman's approach and my own proposal.

1.3 Environment and Body Complexity Thesis

In the literature we can find ideas very similar to the proposal in this paper. In his seminal work on the evolution of cognition, Godfrey-Smith (1996) defended the environmental complexity thesis:

> *"Environmental complexity thesis* (ECT): The function of cognition (and of a range of protocognitive capacities) is to

> enable an agent to deal with environmental complexity." (Godfrey-Smith, 2002, p.135).

We can add that *function* is understood here as "the effect or capacity […] responsible for […] success under a regime of natural selection" (Godfrey-Smith, 2002, p.135). And *cognition* is "a collection of capacities which, in combination, allow organisms to archive various kinds of adaptive coordination between their actions and the world". (Godfrey-Smith, 2002, p.135). Additionally, the environment is not only natural but also social. Therefore, such a cognitive system must also deal with the complex behavior of other living creatures.

Keijzer and Arnellos (2017) describe this view on the evolution of cognition as externalist, where it is shaped by environmental factors to which it is adapted. In response to Godfrey-Smith, they propose a more internalist approach, where not only environmental but also body complexity is important, especially when it comes to multicellular organisms with complex active bodies (see: Trestman, 2013). In these organisms, they see the importance of not only input-output interaction between an organism and its environment but also the internal coordination of internal activity in complex multicellular systems (Keijzer, Arnellos, 2017). For this purpose, the authors propose the concept of the animal sensorimotor organization [ASMO]. They accentuate the "importance of the (internal) multicellular organization as a precondition for the macroscopic environment by animals to become accessible for these animals" (Keijzer, Arnellos, 2017). And it is important that ASMO "fulfils criteria for a minimal cognition" (Keijzer, Arnellos, 2017), and is compatible with our considerations from part (1.1). For them ASMO includes:

> 1. a multicellular body, constituting an 'inner space' or domain, which is differentiated from the body's 'outer space' or environment.
> 2. the presence of contractile epithelia.
> 3. complex, standardized body architectures.
> 4. sensitivity to tension and stress at the level of (intra) cellular processes.
> 5. reversible, contraction-based changes in body-shape. (Keijzer, Arnellos, p.2017).

These conditions are really similar to the role of the body that I describe as constraining conditions (1.2.1). The complex (multicellular) body system must learn its own properties to act and perceive. As we see, cognitive processing is here described in internalist terms (but not only internalist). To be able to cope with changing environmental problems, the system must first be able to coordinate its own stable and changing properties. So, the body is here not only something that enables an animal to perform particular actions—particular ways of dealing with environmental problems. The body, its complexity and coordination is also a problem which must be solved in order to cope with environmental problems. Therefore, even if there are differences between the external and internal (bodily) environments, the animal must coordinate both. Therefore, according to our initial considerations, we should notice that if we consider embodied cognition as the role of cognition for a particular body, we should think of not only the issue of environmental complexity but also of bodily complexity. This leads us to more detailed remarks on cognition and embodiment.

1.3 Toward internal processing in bodily cognitive systems

In the previous parts of this paper, I proposed that in embodied cognition cognition is construed by some kind of a minimal cluster, mainly embedded in sensorimotor coordination problem-solving processes. The body is here described as a whole organism, inducing an important tradeoff between the central and peripheral systems. Then I showed that the brain is not self-sufficient. I finished with remarks on one of the most interesting views on embodied cognition.

From this, we can see that any view of embodied cognition should include an account of internal, cognitive processing[2]. Even if it extends to some body parts going beyond the central nervous system or even some morphological and dynamic properties of the physical body. This processing is, then, highly integrated with the functioning of the whole body and solves the problem raised by body features. It also

[2] It is important to show that the model of cognitive processing in question is appropriate for a system with specific bodily features, so that it is a model of embodied cognitive processing.

exploits some of them to solve some of the problems raised by the environment.

Now, we can look in more detail at the internal processing itself. Here I defend the embodied theory of cognition and show that it is necessary to propose a special approach to internal processing. I will start my consideration by discussing one of the most interesting approaches to this processing. After presenting some of the limitations of this proposal I will develop an alternative approach.

2. Goldman on B-codes and embodied cognition

It is not easy to say what a good theory of internal cognitive processing in embodied cognition should look like. Here I choose one—even if it is not the most widely accepted, it is undoubtedly one of the most interesting proposals: Alvin Goldman's moderate approach to embodied cognition and his idea of B-codes, which embody cognition.

2.1 Moderate approach to embodied cognition

Goldman in his papers (2012; 2014) makes a distinction between a question of the embodiment of cognition in general and the embodiment of a particular cognitive token or exemplar. This distinction seems to be innocent but it is not.[3] I believe that most of the more philosophically oriented research is about a general type of embodiment of cognition (see: Shapiro, 2004; Wilson, Foglia, 2011), but most of psychological work is related to a token or exemplar type of embodiment of cognition. In this context, Goldman (2012; 2014) is an interesting exception because he is interested in embodied tokens or exemplars. This makes his approach more compatible with psychological than philosophical works on embodiment. As already mentioned, in the context of Shapiro's (2010) distinction regarding the relation between embodied and traditional views on cognition, Goldman proposes a moderate approach of embodiment, which is in line with traditional research, and refuses the need of any replacement.

[3] I think it is right to point out similarities between the distinction, present in works about consciousness, between creature consciousness and state consciousness. In this case creature consciousness is analogous to embodied cognition in general, and state consciousness is analogous to embodied tokens or exemplars.

Surprisingly, Goldman also argues that if we describe embodied cognition as a role of the physical body in cognition, we will trivialize this idea. He agrees that when we close our eyes[4] it has an impact on seeing, but says that this is trivial and we cannot base our research on such influence of the body on cognition. I believe and try to show in this paper that, on the one hand, it's impossible to reduce the role of the physical body in cognition to closing eyes or putting fingers in the ears, on the other, that there are non-trivial accounts of the role of the physical body in cognition.

2.2 On B-codes and their re-use

Goldman's approach is based on two ideas. The first one is the idea of the bodily codes or bodily formats. This idea comes from the paper co-authored with Frederique de Vignemont (Goldman, de Vignemont, 2009). The second one is the idea of "reuse", borrowed from the works of Michael Anderson.

The most recent form of Goldman's definition of embodied cognition is as follows:

> Cognition (token) C is a specimen of embodied cognition if and only if C uses some (internal) bodily format[5] to help execute a cognitive task (whenever the task may be) (Goldman, 2014, p.102).

To understand this definition, we need to understand the B-codes and how a system uses these formats to "execute cognitive tasks".

---

4 This example, taken from Goldman papers (2012; and with de Vignemont, 2009), only seems to be trivial but is really interesting. A system with eyes which can be closed could have eyes built of a more fragile and sensitive material, and they could also simply have bigger eyes. A system able to close eyes should have the ability to rapidly update information, taking into account the difference between signals before and after closing eyes. Such a system should integrate visual information with tactile or proprioceptive information for smooth movement coordination in short periods without visual information. Therefore, the fact that I close my eyes is not fundamental for embodied cognition. However, the fact that our eyes can close and open possibly has a big influence on the way we process visual information.

5 Although, Goldman uses the terms *B-codes* and *B-formats* interchangeably, I only use the term *B-code*.

2.2.1 B-codes

In his earlier paper, Goldman (2012) writes that code is something which is "language-like, […] has a distinctive vocabulary, syntax, and a set of computational procedures" (Goldman, 2012, p.73). In his proposal, every sense modality has its own code, and some even have many codes, as in the case of visual perception for action and recognition (Milner, Goodale, 1995). Bodily formats are here described as formats which "in the mind/brain represent states of the subject's own body, indeed, represent them from an internal perspective" (Goldman, 2012, p.73). This is interpreted to mean that bodily formats represent the body through interception, proprioception, and by other somatosensory modalities. However, Goldman describes it neither in detail, nor in terms of syntax, nor computational procedures. We can only say that the syntax and procedures are somatosensory-specific. So we can assume that bodily formats are, to put it widely, somatosensory, internal body representations, primarily involved in body control and representations.

This idea needs more specification. Even if somatosensory, auditory, and visual areas differ from each other and have a distinctive organization (e.g. primary somatosensory cortex is organized somatotopically), at the bottom all neurons work in quite a similar way. For Goldman, codes are distinguished by their connections with separate areas of the brain. B-codes are performed by areas which process information about the body. Of course, there could be many B-codes – there are probably nociceptive, tactile discriminatory and affective codes, and also proprioceptive codes. But what is their nature? In visual perception, the vision-for-action (dorsal) and vision-for-recognition (ventral) streams have differing codes just because one is related to the ventral and the other to the dorsal stream. *Prima facie* it sounds convincing – action coordination and object identification should be executed by distinct computational procedures. However, this difference is one thing, the nature of these codes being the source of the difference is another.

It's possible that this situation is caused by the fact that Goldman is mainly interested in using (actually: reusing) these codes in order to explain cognition, not the brain.

2.2.2 Reuse of B-codes

If Goldman remained interested only in theories of representing and processing information about own body, his approach would be extremely limited. However, he argues that "Embodied cognition is a significant and pervasive sector of human cognition" (Goldman, 2012, p.81). Therefore, he introduces an extension of this theory by adding that: "B-formats are massively redeployed or reused for many other cognitive tasks..." (Goldman, 2012, p. 81).

Based on the results of studies on the activity of the central nervous system, it is argued that a specific type of cognitive activity is embodied (Caramazza et al., 2014; Meteyard et al., 2012; Kubanek, Snyder, 2015). This research indicates that the same areas of the central nervous system are active in the exercise of control tasks as well as in the monitoring of the state of the body and in performing non-related-to-body cognitive tasks. It is not possible to discuss even a small portion of these studies, and, additionally, it doesn't seem to be necessary. We will only use two examples of such research to illustrate the general characteristics of this kind of approach. Goldman (2012; 2014) refers to Pullvemuller's (2005) papers on the connection between language and action, to Glenberg and Kaschak (2002) work on mirror neurons, and to Proffitt and colleagues (2008; 2012) on the role of action and body representations in spatial perception. Because of controversies[6] in his later papers, Goldman (2016) admitted that Proffitt's research cannot be used in his research on B-codes so I will not refer to this research here.

Pulvermuller (2005) relates motor activation to language comprehension. He argues that the motor cortex has somatotopic organization. If language is embodied, then comprehension of action-related words or sentences should also have an effect on somatotopic activation. As Goldman writes (2014, p.96-97), we can observe such activation.

[6] The key point of Proffitt's approach (that is the relationship between the physiological state of the organism and the perception of the steepness of the terrain) was called into question (Durgin et al., 2012; Shaffer et al., 2013). Additionally, Firestone and Scholl (2014) argued that Proffitt's whole concept was based on El Greco fallacy.

> Hearing different sentences involving *lick*, *pick*, and *kick* activated motor areas that control the tongue, the fingers, and the leg, respectively (Goldman, 2014, p.97).

According to Goldman's approach, if B-codes are involved in motor control and the only criterion for the distinction is the brain area, then Pulvermuller's research is an example of the reuse of motoric B-codes for language comprehension.

Goldman's work complements Michael Anderson's (2007; 2010) research on massive redeployment. This approach is based on several theses: (a) in the evolution of the nervous system, old components, if only possible, are reused for new purposes; (b) the same circuits or areas in "different arrangements" i.e., in connection with separate areas are employed for separate tasks, (c) phylogenetically earlier areas are widely connected and more often used for separate task realization. Therefore, areas phylogenetically earlier are the best candidates for being related to B-codes reused later in other not-related-to-body cognitive tasks.

We can now wrap up. The central system uses various B-codes to represent the body, and it reuses these codes to solve problems not related to the body. It seems that the relation to the body is not really so important for Goldman. It is important insofar as we need to distinguish B-codes from other codes. So this is really a weak kind of embodiment, which tells us nothing about dealing with the body and environment complexity.

2.3 From criticism of B-codes to E-codes

Goldman's approach is as interesting as it is controversial. Gallagher points out that in this context there is no real important role for the body itself (Gallagher 2015a; 2015b). In the same vein, Kyselo and di Paolo (2015) write that Goldman's approach is too narrow, and does not include the body's real role in cognition. But the most interesting remarks are in Firestone's paper (2016), who shows that Goldman's use of Anderson's re-use conception is problematic in the case of vision. To be precise, vision can't be embodied in circuits responsible for grasping, because eyes evolved earlier than hands. Goldman (2016) accepted this critique and accepted that his theory doesn't explain the embodiment of vision. Therefore, in Goldman's approach, visual perception is not

embodied. It's a really surprising result, and it's worth remembering, that if vision is in fact embodied, then it is an argument against Goldman's approach.

There is one more critic, important for this paper. At the end of his remarks about Goldman's (2012) approach, Shapiro (2014) directs attention to a very interesting issue. Why can we say that B-codes provide a good account of embodied cognition? More accurately, why are B-codes good in terms of being reused for cognition? Goldman describes B-codes as bodily because they primarily represent the body. But, as Shapiro notices, this is not enough as Goldman doesn't give any reason why they are good for reuse for cognitive purposes. This is, in my opinion, a crucial issue related to Goldman's approach to embodied cognition, and, as Shapiro writes, it "should not be overlooked and is one that places the burden on Goldman to justify his claim that any reuse of B-codes suffices to embody cognition" (Shapiro, 2014, p.87-88). If B-codes are individuated by their primary role of representing the body, and then they are reused because of other reasons, it seems doubtful that we should still maintain that this is embodiment in B-codes.

Further, in this paper I propose a solution which is not dependent on any appeal to the representation of the body.

3. E-codes: Internal processing beyond B-codes

Here I want to sketch some ideas about an alternative to Goldman's view on internal processing in embodied cognition. I call it *E-codes*, because it is coding and processing information embedded in the whole bodily system, and it should be able to give the system the ability to deal with the risk and uncertainty that it must deal with in everyday conditions (see: epigraph at the beginning of this paper). Therefore, contrary to Goldman, I'm interested mainly in an embodiment of cognition in general, not an embodiment of a particular cognitive token or exemplar.

3.1 E-codes: general outline

Embodied cognition need not be a kind of cognition primarily related to the body or about the body. It is essential to consider two types of properties of E-codes: body-specific and body-general. Body-specific properties of E-code are shaped by particular properties of the body. Even if in almost all living organisms their building blocks are quite similar, their structure and organization are quite different. Systems

different in size, morphology, sensors, and effectors solve problems of internal coordination and efficient action in the environment in individual ways. Therefore, cognitive processing differs among them. Body-general properties are also shaped by the body but are related to properties present in all living creatures (energy consumption, dealing with risk and uncertainty). I must point out that the research presented here wasn't developed as research on embodied cognitive processing. It was developed quite independently, but is essential for studies about embodied cognition.

As I argued earlier, cognitive processing is probably, at least partially, extended beyond the central cognitive nervous system (see: Nowakowski, 2015). However, no matter whether this processing extends beyond the central system or not, it should have some properties. In this part of my paper, I try to indicate the kind of properties they should be.

My solution is partially inspired by Keijzer's (2015) research on the evolution of the nervous system as a process of development – a sophistication of a specialized control system. This system is engaged not only in solving problems of interaction with the environment but also internal coordination of neural and muscular tissue activity (see part 1.3). So, we deal here with the problem of efficient action and internal coordination. I describe this as a process of the system "learning" of its properties, possibilities, and constraints. In the context of such processes cognitive systems emerge.

Even if there is not much research on this topic, I can show some body-specific and body-general properties of E-codes:

a. *The laziness of E-codes*: Haselager and colleagues (2008) argue for the lazy brain hypothesis, where the brain in dealing with problems is not searching for the best solution but trying to use the easiest, most accessible, most preferable solutions. Therefore, it's trying to choose the "cheapest", often biased, way to solve the problem. In a similar vein, Clark argues for productive laziness, that cognitive processing should be based on "economic but effective strategies and heuristics" (Clark, 2015, p.244).

b. *Organization and robustness of E-codes*: Our considerations are related to the possible evolution of the whole bodily system. Even though I don't accept Goldman's approach in its entirety, I assume that Anderson's idea of reuse is compatible with E-codes. The evolved system

reuses in any way available subsystems developed earlier. This can be connected to the possible nested organization of a nervous, cognitive system (Bolt et al., 2017). To some degree, we can connect this proposal to a more general idea of degeneracy (more than one subsystem serves a particular function) and redundancy (one subsystem serves more than one function). Such an organization of a cognitive system can increase its robustness and effectiveness[7].

c. *Cost effectiveness of E-codes*: Laughlin (2001) and Niven (2016) argue that energy consumption by the nervous system is a relevant constraint on information processing by the brain. Therefore, brain size, number of connections between neurons, and tradeoffs in processing between the central and peripheral systems are determined by energy consumption. E-codes should be organized in the most energetically economical available way for efficient, fast signaling and minimization of energy consumption at the same time. Wang and Clandinin argue that wiring economy is a significant determinant of nervous system layout (Wang, Clandinin, 2016, p.R1101)

d. *Prospectivness of E-codes*: We can also say, in the context of the motto of this paper, that such bodily systems are almost constantly exposed to the risk of being cheated, injured, or even dying. They must constantly anticipate possible changes in the environment and in their internal milieu. Of course, every system should be anticipatory to some degree. This property can be connected with contemporary works on integration between embodied cognition and predictive processing (see: de Bruin, Michael, 2017; Allen, Friston, 2016; Bruineberg, Kiverstein, Rietveld, 2016; Burr, Jones, 2016). However, there are still controversies about the nature of this integration.

And we can describe some body-specific properties of E-codes:

e. *E-codes and body size and shape*: Organisms of varying sizes and motor flexibility need individual control systems and individual computational procedures (see: Hooper, 2012), various systems to differing degrees offload control on dynamical and mechanical properties of the controlled system.

7 These properties can be increased by balanced (excitation/inhibitory) activation of network and top-down feedback (see: Denève et al., 2017)

f. *E-codes and sensorimotor specificity*: Organisms with individual sensors need individual solutions for effective processing of available sensory information (see: MacIver, 2009). Each individual visual system will need separate kinds of internal processing. An octopus with human eyes will be blind, but for humans seeing with a mantis shrimp eye will be computationally intractable (see: Nowakowski, 2017)

g. *E-codes and various solutions for general problems*: In reference to point 3, we can say that in individual organisms (e.g. with individual body size/brain size ratio), individual solutions for frugal processing are needed.

As we can see, we don't refer to representing or experiencing the body. We don't say what bodies mostly do, only what cognition does for the body, and how it is shaped by the body. But I believe this is the most convincing view on embodied cognition. In our proposal, we describe the embodiment of cognition as an element of the emergence of cognition in the process of effectively coping with the body and with environment complexity.

We can also answer the questions posed in the introduction. It's highly unlikely that B-codes, as described by Goldman, are examples of E-codes. However, if they have to be useful, they should have the properties of E-codes.

Conclusion

Embodied cognition is currently facing problems (Goldinger et al., 2016), so we should search for conceptualizations that are more consistent with the embodiment thesis but that are also consistent with the empirical data. I hope the E-coding approach presented here gives such an opportunity. However, I believe it needs more comparative meta-analysis and computational modeling than psychological experiments, because if cognition is embodied in the way described in this paper, we should observe the correlation between various body morphologies and the various kinds of cognitive processing employed in problem-solving.

If the solution proposed here is correct, it gives the opportunity to develop an account of the "embodied cognitive architectures". We should not forget that embodied cognition is a theory of cognition, not of the body. Cognition in beings *able to take risk,* [...] *and* [beings] *exposed to the unknown.*

Acknowledgments
The work on this paper was funded from National Science Centre research grant under the decision DEC-2014/14/E/HS1/00803. The author wishes to thank Marcin Milkowski for his advice and invaluable support and Anna Karczmarczyk for helpful comments.

REFERENCES

Aizawa, K. (2015a). Cognition and behavior. *Synthese*, 1-20. doi:10.1007/s11229-014-0645-5

Aizawa, K. (2015b). What is this cognition that is supposed to be embodied?. *Philosophical Psychology*, *28*(6), 755-775.

Aizawa, K. (2014). The enactivist revolution. *AVANT. Pismo Awangardy Filozoficzno-Naukowej*, (2), 19-42.

Allen, M., Friston, K. J. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, 1-24.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, *33*(4), 245-266.

Anderson, M. L. (2007). Massive redeployment, exaptation, and the functional integration of cognitive operations. *Synthese*, *159*(3), 329-345.

Bolt, T., Nomi, J. S., Yeo, B. T., Uddin, L. Q. (2017). Data-Driven Extraction of a Nested Model of Human Brain Function. *Journal of Neuroscience*, *37*(30), 7263-7277.

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. https://doi.org/10.1007/s11229-016-1239-1

Buckner, C. (2015). A property cluster theory of cognition. *Philosophical Psychology*, *28*(3), 307-336.

Burr, C., Jones, M. (2016). The body as laboratory: Prediction-error minimization, embodiment, and representation. *Philosophical Psychology*, *29*(4), 586-600.

Caramazza, A., Anzellotti, S., Strnad, L., Lingnau, A. (2014). Embodied cognition and mirror neurons: a critical assessment. *Annu. Rev. Neurosci.* 37,1–15.doi:10.1146/annurev-neuro-071013-013950

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science *Behavioral and Brain Sciences*, *36*(3), 181-204.

de Bruin, L., Michael, J. (2017). Prediction error minimization: Implications for embodied cognition and the extended mind hypothesis. *Brain and cognition*, *112*, 58-63.

Denève, S., Alemi, A., Bourdoukan, R. (2017). The brain as an efficient and robust adaptive learner. *Neuron*, *94*(5), 969-977.

Depraz N. (2005). Radical embodiment, in. H. de Preester, V. Knockaert (ed.), *Body image and body schema: interdisciplinary perspectives on the body* (173 – 186), Amsterdam, Philadelphia: John Benjamins Publishing Company.

Durgin, F. H., Klein, B., Spiegel, A., Strawser, C. J., Williams, M. (2012). The social psychology of perception experiments: Hills, backpacks, glucose, and the problem of generalizability. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(6), 1582.

Firestone, C. (2016) Embodiment in Perception. in. B. P. McLaughlin and H. Kornblith (eds.) *Goldman and His Critics* (318-334), John Wiley & Sons, Inc.

Firestone, C., & Scholl, B. J. (2014). "Top-down" effects where none should be found: The El Greco fallacy in perception research. *Psychological science*, *25*(1), 38-46.

Fuchs, T. (2011). The Brain – A Mediating Organ. *Journal of Consciousness Studies*, *18*(7-8), 196-221.

Gallagher, S. (2015a). How embodied cognition is being disembodied. *The Philosophers' Magazine*, (68), 96-102.

Gallagher, S. (2015b). Reuse and body-formatted representations in simulation theory. *Cognitive Systems Research*, *34*, 35-43.

Gallagher, S., Hutto, D. D., Slaby, J., Cole, J. (2013). The brain as part of an enactive system. *Behavioral and Brain Sciences*, *36*(4), 421-422.

Glenberg, A.M., M.F. Kaschak. (2002). Grounding language in action. *Psychonomic Bulletin and Review* 9, 558–565.

Godfrey-Smith, P. (2002). Environmental complexity, signal detection, and the evolution of cognition, in. M. Bekoff, C. Allen, G.M. Burghardt (eds.) *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition,* (135-142) Cambridge: The MIT Press.

Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge: University Press.

Goldinger, S. D., Papesh, M. H., Barnhart, A. S., Hansen, W. A., Hout, M. C. (2016). The poverty of embodied cognition. *Psychonomic bulletin & review*, *23*(4), 959-978.

Goldman, A. I. (2016). Response to Firestone. in. B. P. McLaughlin and H. Kornblith (eds.) *Goldman and His Critics* (335-336), John Wiley & Sons, Inc.

Goldman, A. I. (2014). The bodily formats approach to embodied cognition, in. U. Kriegel (ed.) *Current Controversies in Philosophy of Mind*, (91-108), Routledge.

Goldman, A. I. (2012). A moderate approach to embodied cognitive science. *Review of Philosophy and Psychology*, *3*(1), 71-88.

Goldman, A., de Vignemont, F. (2009). Is social cognition embodied?. *Trends in cognitive sciences*, *13*(4), 154-159.

Haselager, P., van Dijk, J., & van Rooij, I. (2008). A lazy brain? Embodied embedded cognition and cognitive neuroscience. in. P. Calvo, T. Gomila (eds.) *Handbook of cognitive science: An embodied approach*, ed. (273-287), Elsevier Science.

Hooper S.L. (2012). Body size and the neural control of movement, *Current Biology*, 22(9), R318-R322.

Jacob, P. (2012). Embodying the mind by extending it. *Review of Philosophy and Psychology*, *3*(1), 33-51.

Jékely, G., Keijzer, F., Godfrey-Smith, P. (2015). An option space for early neural evolution. *Phil. Trans. R. Soc. B*, *370*(1684), 20150181.

Keijzer, F. (2015). Moving and sensing without input and output: early nervous systems and the origins of the animal sensorimotor organization. *Biology & Philosophy*, *30*(3), 311-331.

Keijzer, F. (2003). Making decisions does not suffice for minimal cognition. *Adaptive Behavior*, *11*(4), 266-269.

Keijzer, F., Arnellos, A. (2017). The animal sensorimotor organization: a challenge for the environmental complexity thesis. *Biology & Philosophy*, 1-21. https://doi.org/10.1007/s10539-017-9565-3

Kubanek, J., Snyder, L.H. (2015). Reward-based decision signals in parietal cortex are partially embodied. *J.Neurosci.* 35,4869–4881.doi: 10.1523/JNEUROSCI.4618-14.2015

Kyselo, M., Di Paolo, E. (2015). Locked-in syndrome: a challenge for embodied cognitive science. *Phenomenology and the cognitive sciences*, *14*(3), 517-542.

Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. *Current opinion in neurobiology*, *11*(4), 475-480.

MacIver, M. A. (2009). Neuroethology: From Morphological Computation to Planning. in. P. Robbins, M. Aydede (ed.) *The Cambridge Handbook of Situated Cognition* (480-504), Cambridge: Cambridge University Press.

Meteyard, L., Cuadrado, S.R., Bahrami, B., Vigliocco, G. (2012). Coming of age: are view of embodiment and the neuroscience of semantics. *Cortex* 48, 788–804.doi:10.1016/j.cortex.2010.11.002

Milner, A. D., Goodale, M. A. (1995). The visual brain in action, *Oxford: Oxford University Press*.

Miłkowski, M. (2016). Models of Environment. In Frantz, R., Marsh, L. (eds.) *Minds, Models and Milieux* (227-238). Palgrave Macmillan UK.

Niven, J. E. (2016). Neuronal energy consumption: biophysics, efficiency and evolution. *Current opinion in neurobiology*, *41*, 129-135.

Nowakowski P.R. (2015). Commentary: The Embodied Brain: Towards a Radical Embodied Cognitive Neuroscience. *Front. Hum. Neurosci*. 9:623. doi: 10.3389/fnhum.2015.00623

Nowakowski P.R. (2017). Bodily processing: the role of morphological computation, *Entropy, 19*(7), 295; doi:10.3390/e19070295

Proffitt, D.R. (2008). An action-specific approach to spatial perception. in. R.L. Katzky, B. MacWhinney, M. Behrmann (eds.) *Embodiment, ego-space, and action*, (177-200), Psychology Press

Proffitt, D.R., Linkenauger. S.A. (2012). Perception viewed as a phenotypic expression. in.W. Prinz, M. Beisert, A. Herwig. (eds.) *Action Science. Foundations of an Emerging Discipline*. (171-198) Cambridge: MIT press.

Pulvermuller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience* 6, 576–582.

Rupert, R. D. (2016). Embodied Functionalism and Inner Complexity: Simon's Twenty-first Century Mind. In Frantz, R., Marsh, L. (eds.) *Minds, Models and Milieux* (7-33). Palgrave Macmillan UK.

Shapiro, L. A. (2014). When is Cognition Embodied, in. Uriah Kriegel (ed.) *Current Controversies in Philosophy of Mind* (73-90) Routledge,

Shapiro, L. A. (2010). *Embodied cognition,* Routlege.

Shapiro, L. A. (2004). *The mind incarnate*. Cambridge: MIT press.

Shaffer, D. M., McManama, E., Swank, C., Durgin, F. H. (2013). Sugar and space? Not the case: Effects of low blood glucose on slant estimation are mediated by beliefs. *I-Perception*, *4*(3), 147-155.

Trestman, M. (2013). The Cambrian explosion and the origins of embodied cognition. *Biological Theory*, *8*(1), 80-92.

Van Duijn, M., Keijzer, F., Franken, D. (2006). Principles of minimal cognition: Casting cognition as sensorimotor coordination. *Adaptive Behavior*, *14*(2), 157-170.

Wang, I. E., Clandinin, T. R. (2016). The influence of wiring economy on nervous system evolution. *Current Biology*, *26*(20), R1101-R1108.

Wilson, R. A. (2010). Extended vision, in. Gangopadhyay, N., Madary, M., Spicer, F.(ed.) *Perception, action and consciousness*, (277-290). Oxford, New York: Oxford University Press.

Wilson, R. A. (1994). Wide computationalism. *Mind*, *103*(411), 351-372.

Wilson, R.A., Foglia, L. (2011/2017). Embodied Cognition, in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy,* URL = <https://plato.stanford.edu/archives/spr2017/entries/embodied-cognition/>.

Wilson A.D., Golonka, S. (2013). Embodied cognition is not what you think it is. *Front. Psychology* **4**:58. doi: 10.3389/fpsyg.2013.00058

ABSTRACT

EMBODIED COGNITION: LOOKING INWARD

The body is a highly complex, coordinated system engaged in coping with many environmental problems. It can be considered as some sort of opportunity or obstacle, with which internal processing must deal. Internal processing must take into account the possibilities and limitations of the particular body. In other words, even if the body is not involved in the realization of some cognitive explicit task, it is not a neutral factor of our understanding of why a system solves a task in one way or another. Therefore, when conducting research on embodiment and the body's cognitive system we should not neglect internal, cognitive processing.

I appeal to Goldman's research on embodied cognition to sketch the broader framework for internal processing in embodied cognition. I believe that even if we don't accept Goldman's approach as the viable proposal for embodied cognition in general, it's a quite natural starting point for our analysis. Goldman (2012; 2014, and with de Vignemont 2009) argue for the essential role of the bodily formats or bodily codes (respectively: B-formats and B-codes) in embodied cognition. B-codes are here described as the processing of regions or sub-regions of the central nervous system. They are primarily employed for body control or monitoring, and reused for cognitive tasks. Beyond doubt, this conception provides an excellent starting point for analyzing the internal (mostly neural) processing in cases of embodied cognition.

At the end of this paper, I will argue that the embodiment of cognition needs a conceptual twist. Following Keijzer's (2015) interest in the evolution of the nervous system, and the minimal forms of cognition, I argue that in investigating embodied cognition, we should investigate the role played by cognitive processing for specific kinds of organisms, meaning organisms with a body of a particular morphology (size, shape, kinds, and distribution of sensors and effectors). Doing that, I refer to some conceptual and empirical considerations. I will also try to show that research on embodied cognition is still not sufficiently anchored in evolutionary and comparative studies on cognition, nor on the nervous system and body morphology. Bigger reliance on these kinds of studies,

will make it make possible to gain a deeper understanding of internal processing in embodied cognition.

KEYWORDS: embodied cognition; bodily cognitive system; internal and cognitive processing; B-codes; E-codes

HYBRIS nr 38 (2017)
ISSN: 1689-4286

PAWEŁ GŁADZIEJEWSKI
INSTITUTE OF PHILOSOPHY AND SOCIOLOGY
POLISH ACADEMY OF SCIENCES

JUST HOW CONSERVATIVE IS CONSERVATIVE PREDICTIVE PROCESSING?[1]

Introduction

According to the Predictive Processing (PP) framework, perception, action, and perhaps a large portion of cognition are underpinned by a mechanism of prediction error minimization. On this view, the central nervous system builds a hierarchical generative model whose job is to recapitulate the causal structure of the environment. The model generates a cascade of 'mock' predictions about incoming sensory stimuli. These predictions are matched against actual input and revised to minimize the discrepancy between the way the sensory organs are stimulated and the way they are predicted to be stimulated. What gets propagated up the hierarchy is just the prediction error signal that signifies the divergence between the two. Each level of the hierarchy minimizes the error only relative to a level directly below. The gain on the prediction error signal is mediated by precision estimations, so that, depending on the variance of the sensory signal, the processing can be modulated to rely more on the input or the internal dynamics ('prior knowledge') of the system. Perception on this view is a matter of minimizing the error by matching the internal estimates ('hypotheses') to actual environmental causes of the sensory signal. Action

[1] Work on this paper was supported by the Polish National Science Centre FUGA 3 grant (UMO-2014/12/S/ HS1/00343).

is a matter of intervening on the environment to match its state to internal estimates so that the prediction error is minimized (for reviews, see: Clark, 2013, 2016b; Hohwy, 2013; Wiese & Metzinger, 2017). And cognition can be hypothesized to result from an off-line activation of the same predictive machinery primarily involved in perception and action (see Pezzulo, 2017).

PP is surrounded by an aura of revolution, as many see in it an extremely ambitious framework that promises to provide a long-awaited (at least by some) theoretical unification for the sciences of cognition. How this supposed revolution fits into existing debates about the nature of cognition is now hotly debated. PP was initially construed in a manner that dovetails with traditional approaches in cognitive science, i.e. ones that see cognition as matter of inferential, exclusively intracranial processes involving richly structured representational states (Hohwy, 2013, in press). Following Clark (2015, 2016b), I will call this interpretation of PP 'conservative'. However, recently a number of researchers have argued that construing PP in conservative terms is mistaken. These authors opt for a 'radical' reading of PP, one that marries the framework with the idea that cognition is completely or largely non-representational as well as body- and environment-involving (Allen & Friston, 2016; Bruineberg, Kiverstein & Rietveld, 2016; Clark, 2016a, 2016b; Hutto, 2017; Orlandi, 2016, 2017). Such views situate PP firmly within the 4E approaches to understanding and studying cognition.

It seems that the literature is now shifting toward this latter, radical reading of PP. Perhaps one major reason behind this is the recognition that the PP framework finds proper theoretical home within the larger context of the Free Energy Principle (FEP; see e.g. Friston, 2010, 2013; Friston & Stephan, 2007). The FEP states that to avoid circumstances with high surprisal (i.e. ones that endanger the organism's homeostatic integrity and are unlikely given its phenotype), living creatures minimize the information-theoretic quantity of free energy. The FEP comes from a theoretical biology and applies to all, even single-cell organisms. Because, under Gaussian assumptions, long-term prediction error is equivalent to free energy, there is a tight connection between FEP and PP. PP naturally emerges as a theory of how the central nervous system, in some species, enables organisms to self-organize by avoiding surprising states. Perhaps

PP provides a sketch of a causal mechanism through which living creatures implement the FEP (Klein, 2016). The exact nature of the connection between FEP and PP is beyond the scope of the paper. I take it that although FEP puts crucial constraints on our understanding of PP (this will become apparent in the discussion to come), the two can be considered as distinct to a degree. One is a theory of life, the other is a theory of cognitive architecture tightly connected to the first. In this paper, I focus on the latter.

The aim of the present paper is to revisit the conservative construal of PP, as it is not entirely clear what this approach to understanding the framework is committed to exactly. It is all too easy to treat the conservative approach as naively attached to an outdated, overly intellectualist and internalist view of cognition. I aim to review, clarify, and disentangle the conservative commitments of PP. I take these commitments to be distinct from each other and at least partially independent. I propose that these commitments are threefold: (1) the commitment to representationalism; (2) the commitment to the notion of inference as subserving perception and action; and (3) the commitment to internalism, where internalism means that the constitutive basis of cognition does not extend beyond the central nervous system. I want to investigate and interpret each of those commitments in a way that is both grounded in PP and charitable towards proponents of the conservative approach. The discussion to follow will show that whatever genuine conservatism can be found in PP, it is as ecumenical towards 4E approaches as conservatism gets (this amounts to a intermediate, moderate position, not unlike the one proposed in: Dolega, 2017). This paper is largely a review which aims to group ideas already scattered throughout the literature and show how they fit together.

I start (in Section 1) by addressing the role that the notion of representation plays in PP. I argue that this notion can be interpreted in a weak (pragmatist) or strong (strictly realist) way. I claim that even realistically construed, representations as postulated by PP are largely within the spirit of the 4E approaches. In Section 2, I argue that PP makes use of a liberal, and yet non-trivial notion of inference. This sort of inferentialism boils down to the claim that the transitions between representational states postulated by PP are under internal control and truth-preserving (they approximately follow a truth-preserving rule). In

Section 3, I argue that PP's pretensions to internalism are not justified by the conceptual resources of the predictive framework itself. In particular, the notion of a Markov blanket is not enough to justify the commitment to internalism. I discuss how PP relates to some other, internalism-friendly ways of delineating the boundaries of mind already present in the literature.

1. The commitment to representationalism

1.1. Weak and strong representationalism of PP

Perhaps the most obvious motivation to treat PP as committed to representational states stems from the fact that the framework conceptualizes perception in terms of Bayesian inference. Minimizing the prediction error can be treated as equivalent to maximizing the posterior probability of hypotheses about the causes of the incoming sensory signal. When looked at this way, PP is simply filled with semantic notions. The perceptual system comes up with '*hypotheses' about* distal states of environment, using *'beliefs' about* which distal causes are most likely (priors) and *about* what sort of sensory 'evidence' is to be expected given some hypothesis (prior likelihoods). These hypotheses and prior beliefs are semantically evaluable: they can go wrong in the sense of *misrepresenting* the way things are. This all should not come as a surprise, as in any Bayesian theory of perception, perceptual states are individuated by their representational relations to the environment (Rescorla 2013).

However, the mere fact that semantic notions are at use does not necessarily mean a win by default for a proponent of a conservative reading of PP. There are in fact two significantly distinct ways to understand PP's commitment to representationalism. On what we can call a weak reading, the representational notions at play merely serve as what Frances Egan calls 'intentional gloss' (Egan, 2010, 2014; for proposals that explicitly interpret PP's commitment to representationalism by invoking Egan's account of conent, see: Downey, 2017; Wiese, 2017).[2] On Egan's account of content and its role in cognitive science, to make sense of physical transactions within a

[2] Note that (Wiese 2017) does not endorse Egan's pragmatism about content and is in many respects closer to a strong reading of PP's representationalism, which will be discussed in the main text.

given (computational) system of interest, its internal structures and states are mapped onto abstract mathematical entities (like numeric values). The attribution of 'mathematical contents' enables the researchers to make sense of the computations (e.g. the operation of addition) that the system in question performs. However, this is not enough to get a full understating of the system engaged in some environment-specific cognitive task. To explain how computing some function contributes to the exercise of a cognitive capacity, 'cognitive' contents must be ascribed, i.e. contents that relate parts of the internal machinery to parts of the task-specific environment. According to weak reading of the representational commitment, this is exactly the case with the sematic notions at use in PP. These contents are ascribed to the error-minimizing computational machinery to get an understanding of how it is related to the environment, a feat that is hardly achievable with purely physical and computational description.

Now, the important thing to take from this is that under this weak interpretation, any content to be found in perceptual states postulated by PP is of derived nature. Intentional properties (cognitive contents) are ascribed to the internal machinery for purely pragmatic reasons. That is, the internal states do not have cognitive contents intrinsically or essentially, but purely in virtue of interpretative acts on part of the researchers engaged in explaining cognitive functions. Thus construed, content is not a causally efficacious property of 'hypotheses' or 'prior beliefs', but may be rather seen as nothing more than a useful fiction (Downey, 2017; for a discussion of fictionalism about representation, see Sprevak, 2013). Overall, this sort of view renders PP representational in such a minimal sense that not many proponents of the 4E approaches would presumably be moved by it. After all, on this weak reading, what we are dealing with is simply a representational gloss on a non-representational mechanism. The representational vocabulary may be of crucial heuristic value, but cognition as such turns out contentless.

Still, there is a far stronger way to interpret PP's commitment to representationalism. On this reading, PP postulates a rich set of states with real, causally efficacious representational content. The justification for such a view comes from a close inspection of the role played in PP's overall computational machinery by the generative model. The generative model is

supposed to 'recapitulate' the causal structure of the environment and send a top-down stream of multi-level, cascading sensory predictions. There are strong reasons to regard the generative model as contentful and engaged in a nontrivially representational role (for more detailed and closely related discussions, see: Gładziejewski, 2016; Kiefer & Hohwy, 2017; Wiese, 2017; Williams, 2017). First, it generates, in perceptual inference, estimates of the environment which guide cognitive system's practical engagements with the environment. It is action-guiding. Second, the model's ability to play this function is dependent on how well the functional relationships between encoded variables resemble the causal structure of the environment. The degree of structural match between the model and the environment is causally relevant to a degree in which the model is effective at enabling adaptive, self-maintaining actions (see Gładziejewski & Miłkowski, 2017). This way, content becomes the fuel of practical success, not just a matter of passively mirroring the environment. Third, the model performs a largely endogenously-controlled, predictive simulation. It exhibits at least some degree of detachment or independence from current sensory stimulation. It could be argued that the simulations in question can be run purely off-line, i.e. outside of any direct engagements with the environment (Pezzulo, 2017). Fourth, insofar as the model undergoes correction in light of the prediction error, it can be said to be capable of detecting cases when its representations are inaccurate. More precisely, the Kullbach–Leibler divergence between true posterior and recognition (model-based) probability distributions can be understood as a sort of measure of misrepresentation (Kiefer & Hohwy, 2017). The lesson, then, is that the generative model constitutes an action-guiding, detachable structural representation, capable of detectable representational error. This is a robust and metaphysically realist incarnation of representationalism, arguably immune to recent trivializing arguments against representation (see Gładziejewski, 2015, 2016; Gładziejewski & Miłkowski, 2017).

1.2. Strong representationalism about PP: how conservative?

Let us focus further on PP's strong representationalism, as this is what proponents of 4E approaches would presumably take issue with. It could be suggested that by invoking the concept of an internal model, conservative

rendering of PP construes representations involved in perception as action-neutral, disembodied inner replicas or reconstructions of the world (Clark, 2015, 2016b). On closer inspection, this sort of assessment turns out unfair towards conservatism. In fact, as far as robust and metaphysically realist representationalism goes, the (strong) notion of representation in PP is very much compatible with the spirit of 4E approaches.[3] There are four reasons to see PP's commitment to strong representationalism as not-so-conservative after all.

First, note that PP postulates a complex processing architecture subserving the process of minimizing the prediction error. The generative model is just a part, albeit important, of this larger architecture. It is entirely possible that this scheme includes both representational and nonrepresentational aspects or parts. Even strong commitment to representationalism in PP does not have to entail a view on which *all* there is to cognitive processing is representation-munching. In addition to the generative model, PP comes with at least three other posits: (1) the sensory signal which results from the world affecting the sensory apparatus of an organism, (2) the prediction error signal which is propagated bottom-up, and (3) precision estimators which regulate the gain on the prediction error signal. For each of those posits, we may ask whether its functioning is representational in nature. Although a case could be made that precision estimators are representational (Wiese, 2016), the same may not apply to the sensory signal. The latter acts as a mere causal mediator incapable of representational error (Gładziejewski, 2017). And there is still a further question of whether the bottom-up error signals earn a representational reading (Orlandi, 2016 can be read as providing a negative verdict here). The point is that PP does not come with wholesale representationalism; there may be purely non-representational structures and processes involved in perception and action control.

Second, even on the strong reading of PP's representationalism, the representations in question are *anything but* action-neutral. Remember that considered in the context of FEP, the process of minimizing the prediction error is merely a way of achieving a pragmatic goal of keeping an organism

[3] That is unless, of course, one is committed to full-blown antirepresentationalism.

within conditions that help maintain it in a far-from-thermodynamic-equilibrium state. This is directly achieved through action, construed in PP as minimizing the prediction error by engaging reflex arcs to quash proprioceptive prediction error. And perception (perceptual inference) is there to provide guidance for action; estimating the causes of the sensory signal functions to enable adaptive engagement with environment. In other words, on the PP view of things, building a structural representation of environment is not an end in itself but a tool of self-maintenance (Williams, 2017). This is in line with those approaches in the literature that try to recast representationalism so that it becomes not an alternative but an ally to 4E approaches (Bickhard, 1999; Rosenberg & Anderson, 2004).

Third, the content of representations postulated by PP is organism-relative and shaped by the organism's embodiment. To see this, PP once again must be considered within the proper context of FEP. Given that perception is ultimately a tool for self-maintentance, the content of the internal models is naturally expected to be strategically selective (Burr & Jones, 2016; Clark, 2013, 2015; Williams, 2017). What is 'reconstruced' in internal models of prediction-error-minimizing-agents are those aspects of the environment which constitute the organism's Umwelt, i.e. the ones which the organism depends on in its practical engagements with the environment. Furthermore, given that one situation can be associated with different suprisals for different types of organisms (what has large surpirisal for a human phenotype may not be surprising for a cod phenotype), it is natural to hypothesize that the content of those models will differ from species to species (Williams, 2017). Also, the organism's body plays a non-trivial role in constraining the contents of generative models. To learn the causal structure of its surroundings, the prediction-error-minimizing-agent needs to intervene on the environment, where those interventions serve as 'experiments' that enable the system to disambiguate between alternative hypotheses. The body plays a crucial role here, as it serves as a reliable, readily-available 'laboratory' (Burr & Jones, 2016). The sort of statistical patterns most readily accessible and learnt are those that the depend on the bodily interactions with environment.

Fourth, consider the question of the vehicles of representations in PP. Here, of particular interest is how PP deals with the idea of detached

representations, that is representations used for off-line cognition instead of for perception or action control. On PP view of things, imagery, counterfactual reasoning, action planning or dreaming could be understood in terms of generative models run is simulation mode – in a way that is fully or partially freed from the "sensory enslavement" of direct interaction with the environment (see e.g. Hobson & Friston, 2012; Pezzulo, 2017; Seth, 2014). Simulations of this sort could generate a cascade of top-down sensory signals, activating levels relatively low within the hierarchy. This way, generative models could run simulations that span multiple levels of the processing hierarchy and bring about patterns of neural activity that resemble to those that accompany perception and action. If this is so, then representational vehicles underlying off-line cognition will not comprise of amodal, body-neutral neural code, but will rather involve neural machinery primarily involved in modality-specific (this includes interoception, see Seth, 2013) on-line cognition. This again connects nicely with what some proponents of the embodied approach have argued for (Barsalou, 1999; Goldman, 2012).

2. The commitment to inference

The second conservative commitment of PP relates to the notion of inference. The motivation for it stems from the idea of the external world as a sort of 'black box' for the skull-bound brain (Clark, 2013; Hohwy, 2013). On this story, to do its job as a controller of action, the brain needs to generate movements that accord with the layout of the organism's immediate surroundings. A real-life snake and a snake-looking cucumber mandate different reaction on part of the agent. However, all that the brain has direct access to are the effects that the external things impinge on the sensory apparatus of the organism. The input is ambiguous, as sensory states are underdetermined by the world: in many realistic circumstances, the sensory effects of a snake and a cucumber may be quite similar. Hence, the task of perception is to recover the *most likely* external causes of the sensory signal – out of a range of some alternatives – so that adaptive action can be initiated. This 'recovery' is construed in terms of an inference under uncertainty. The brain abductively 'infers' environmental causes of the sensory input, that is, it comes up with hypotheses that best explain (given

a larger model of the environment) the sensory patterns by citing their worldly causes. This, of course, places PP within a longer history of thinking about perception in terms of an abductive inference (Gregory, 1980; Helmholtz, 1860/1962).

The idea that PP is in fact committed to inferentialism about perception faces two sorts of criticism. On the one hand, it may be argued that the view presented above gets the 'epistemic' situation of the brain completely wrong. Perception is not underdetermined by sensory stimulation because all the required information is already present in the physical energies affecting the sensorium; and/or because the brain is, in virtue of its wiring, attuned to statistical patterns in the environment to the degree where no disambiguating inference is needed (Anderson, 2017; Orlandi, 2016, 2017). There is no motivation for postulating inference in the first place. I will not address this sort of criticism here, as it seems to be properly aimed not at the *conservative* reading of PP, but the whole PP framework itself. It arguably makes obsolete the very postulate that the brain implements a nesting, hierarchical model engaged in generating top-down sensory prediction. On the other hand, it may be argued that the notion of inference at play in PP it is either trivially liberal or misconstrues what the framework actually postulates (Bruineberg, Kiverstein & Rietveld, 2016). That is, the 'inferences' involved are not genuine inferences or the inferential approach is not justified by what the PP says about the machinery underlying perception, action, and cognition. To address this sort of criticism, I want to first elucidate what 'inference' as postulated in PP amounts to, and then proceed to show that it is neither excessively liberal nor does it get PP wrong (for a similar, in-depth defence of the inferential nature of PP and related computational models of perception, see Kiefer, 2017).

There are three crucial ingredients that make PP genuinely inferential. First, note that the commitment to inference is strongly tied to the commitment to representation. Given that inference constitutively involves transitions between *contentful* states, the former commitment presupposes the latter. In fact, it seems that to treat inference as postulated in PP literally, we should go with the stronger, realist brand of representationalism. Assuming strong representationalism, there *are*

transitions between genuinely contentful states in PP, as the internal hierarchical generative model is changing to keep track of the environment at different time-scales. This amounts to updating an action-guiding, detachable, error-detection-affording structural representation of the environment. Two general transitions involved are (1) revising the current estimate to match the current sensory input; (2) learning through perception, that is, revising the overall structure of the model ('priors') so that the prediction error is better minimized over longer periods. The model goes from one representational state to another by revising, adding, or dropping current hypotheses and long-standing beliefs.

Second, these representational transitions are approximately Bayesian without explicitly representing the Bayes rule. It is reasonable to hypothesize that a system that minimizes prediction error is a system that performs approximate Bayesian inference by maximizing the posterior probability of its model of the environment (see Hohwy, in print; Hohwy, Roepstorff & Friston, 2008; Kiefer & Hohwy, 2017). This means that a system updating its generative model to minimize prediction error is a system that updates its internal estimates of the environment in a way that conforms with Bayes rule. As such, given that Bayesian inference embodies a rational rule for revising one's beliefs or subjective probabilities, perception (and action, see Hohwy, in print) on PP view turns out to conform to a normative principle. Its rationality stems from the fact that Bayesian inference is truth-preserving (for a more detailed discussion, see Kiefer, 2017). And truth-preservation is another constitutive feature of inference.

Third, it seems that a kind of autonomy is implied in truly inferential processes. Suppose that there is succession of events A, B and C and that each of those events produces, in turn, an internal representational state A', B' and C' in some cognitive agent. Suppose that the move from A' to B' to C' conforms to some truth-preserving rule like modus tollens. Because of how the transition between the representational states is completely determined by external events, it does not seem to count as inference. Inference is constitutively an act, a part of agent's cognitive *activity*. Importantly, representational transitions involved in PP meet this criterion of inference. The way that perceptual hypotheses and priors are updated is not a matter of passively registering external states. Rather, it is co-shaped by the

internal states and dynamics of the prediction-error-minimizng system. The perceptual inference and perceptual learning are not completely determined by the driving, sensory signal, but actively shaped and constrained by the system's prior 'knowledge'. So, inference properly counts here as an active, not just reactive process.

Taken together, this amounts to a view of inference as an act of representational change that (approximately) conforms to a truth-preserving rule. *This*, and nothing more, is the sense in which conservative PP is committed to inference. Notably, there may be other considerations in favor of the claim that literal inference is involved in PP. For example, Kiefer (2017) argues that – in line with some influential treatments of inference in philosophical literature – representational transitions in PP (and related frameworks) are such that they increase the overall coherence of representations involved, that is, their consistency and the number of inferential connections between them. Another point might be that because the generative model reduces the prediction error relative to the sensory signal (as caused by the external world), the representational change can be also seen as maximizing the 'empirical adequacy' of the model. Nonetheless, it must be conceded that the sort of inferential processes postulated in PP also *lack* some of the features that characterize many paradigmatic instances of inference. In particular, they are not consciously accessible or goal-directed in the sense of being driven by personal-level intentions. But it is doubtful whether any of those features is *necessary* for a cognitive process to count as inference (see Kiefer, 2017).

As mentioned, the idea that full-blown inference is involved in PP can raise some skepticism. One reason for this stems from a close inspection of the way that the notion of inference is employed in the literature on FEP. As some authors point out (Bruineberg, Kiverstein & Rietveld, 2016), 'inference' as used in the work of Karl Friston (e.g. 2013) boils down to a dynamic coupling between the organism and its environment in which the mutual information between the internal (organismal) and external ('hidden') states is maximized. Because, almost by definition, every organism falls under FEP (to live is to actively avoid surprising and seek unsurprising states), every organism can count 'inferring' the states of the environment in this sense. Furthermore, this notion applies to non-living

coupled systems, for example, to a system composed of two coupled pendulum clocks (Bruineberg, Kiverstein & Rietveld, 2016) There clearly is something misleading about treating bacteria or synchronized clocks as engaged in *literal* inference. This very minimal, relaxed usage of the notion diverges from a more cognitivist sense that most associate with inferentialist view of perception. However, as mentioned at the outset of this paper, we need to be careful to distinguish between PP and FEP. This raises the possibility that the notion of inference at play in PP is different than the one sometimes used in discussions of FEP. And it seems that this is exactly the case. 'Inference' at use in PP is significantly stronger: it entails far more than the coupling of two dynamic systems. It involves an endogenously controlled transition between genuinely representational states that approximately conforms to a truth-preserving rule. Hence, the concerns about trivialization of the notion of inference which can be reasonably raised in the context of FEP do not apply to PP.

Another way to challenge the inferential reading of PP is by trying to show that the processes the framework postulates have features that prevent them from counting as truly inferential. In particular, some authors (Bruineberg, Kiverstein & Rietveld, 2016) point to the fact that traditional inferential theories of perception rely on an analogy between perception and scientific hypothesis testing. But this analogy collapses once we consider PP in the context of FEP. When properly construed, the job of the perceptual system is not to generate representations that 'objectively' capture the environment. Perception is a fundamentally biased sort of hypothesis-testing enterprise:

> If my brain really is a scientist, then it is heavily invested in ensuring the truth of a particular theory, which is the theory that "I am alive". This is a fundamental prior belief that drives all action; namely, I exist and I will gather all the evidence at hand to prove it. It will only make predictions whose confirmation is in line with this hypothesis. It does not give competing hypotheses a fair chance and is extremely biased in the way it interprets the data. It decides on the outcome of an experiment beforehand (my staying alive) and manipulates the experiment until the desired result is reached. If my brain is a scientist, it is a crooked and fraudulent scientist (…) (Bruineberg, Kiverstein & Rietveld, 2016, pp. 14-15).

One might feel tempted to use these considerations as an argument against the involvement of inference in PP. But this criticism would beg the question. Of course, according to PP, the perceptual system is *not* interested in truth for the sake of it. As mentioned before, it is selective in the way it recapitulates the structure of the environment. It is natural to expect that it changes its representational states in a way that is systematically biased toward the overarching aim of keeping the organism in unsurprising states, which sometimes means sacrificing truth or accuracy. Furthermore, it has been forcefully argued on PP view of things, action initiation is based on systematically *misrepresentational* precision estimations (Wiese, 2016). Yet, it is far from clear why the fact that the way the perceptual system works diverges from idealized norms of scientific rationality could prevent the system in question from counting as *inferential*. Because of social factors and cognitive biases, the way *scientists* update their hypotheses in light of evidence sometimes (perhaps often) deviates from idealized norms of scientific rationality. This hardly makes the updating process non-inferential. To generalize, crooked inference is inference nonetheless. And as I take it, conservative rendering of PP (charitably interpreted) is *only* committed to the idea of perception as inference, *not* to an importantly different and stronger claim that perceptual inference functions to uncover truth for the sake of it.

3. The commitment to internalism

The last commitment often associated with conservative construal of PP is to an internalist view of the mind. Here, 'internalism' means a claim that, contrary to extended and (strong incarnations of) embodied views, the constitutive basis for cognition does not go beyond the boundary of the central nervous system. This 'neurocentric' or 'seclusionist' reading of PP is defended by appealing to the notion of a Markov blanket (Hohwy, 2016, 2017, in print). The concept comes from causal network models and refers to nodes of the network such that, given some node X, the state of X is statistically fixed (can be fully predicted) by the states of those nodes. The Markov blanket of X will thus include its neighboring nodes: its 'parents' (proximal nodes that activate X), its 'children' (proximal nodes activated by X) and the parents of its children (Friston, 2013). Now, the point is that

internal sensory and 'active' (motor) states constitute a Markov blanket for a prediction-error-minimizing agent. Less technically, to fully predict how agent's internal states will evolve in time, all that is required is knowledge about its internal dynamics and what happens at the sensorimotor Markov blanket. Assuming that on the PP view of things cognition is prediction-error-minimization, the generative-model-based machinery involved in minimizing the error is situated within the Markov blanket thus construed. This way, the brain and spinal cord emerge as the sole seat of mindedness. Relatedly, this also opens up the possibility of skepticism, whereby an agent can enjoy a rich cognitive life even if it is being fed its sensory states not by the external world (nor does it output its active states to actual body), but rather by a misleading demon.

As noted by the opponents of the conservative reading of PP, this way of defending internalism in PP turns out problematic (Clark, 2016a, 2017; Fabry, 2017). One particularly forceful criticism points out that the concept of a Markov blanket is a technical notion that can be applied to any dynamical system to demarcate it from its environment (Clark, 2017). There will be Markov-blanketed systems *within* the prediction-error-minimizing agent, from single neurons to particular levels within the hierarchical generative model implemented in the brain. In addition, Clark argues that nothing prevents us from postulating Markov-blanketed systems that encompass the (embodied) brain *and* parts of the external, technological environment. That is, a system that comprises the biological agent equipped with technological extensions or interfaces could count as prediction-error-minimizing agent enclosed within a Markov blanket. In fact, Clark (2017) opts for a view that the boundaries of minds change 'metamorphically' through life as technological extensions are added and subtracted.

Assuming there is a nesting hierarchy of Markov-blanketed systems that go both within and outside the brain, natural questions arise. Which Markov blanket is the privileged one when it comes to delineating the mind? And why think that the boundary coincides with the blanket that secludes the central nervous system? In fact, these considerations leave us with three options regarding the idea of a Markov blanket as cognition- or mind-delineating boundary: (1) there is one, stable, unique blanket that delineates cognition and it is the blanket that surrounds the central nervous system;

(2) the boundaries of a cognitive system are enclosed by a Markov blanket that metamorphically changes to include factors that go beyond the central nervous system alone; (3) no Markov blanket serves as a unique, cognition-demarcating one. Only option (1) counts as genuinely conservative. However, the most important lesson is that the technical notion of the Markov blanket *as such* is not enough to decide between these three options (Clark, 2017). This means that the justification for internalist reading of PP, if it is to be found at all, presumably will not come from the conceptual resources of the framework itself.

Internalism turns out to constitute a soft underbelly of conservatism about PP, the one commitment that seems the least justified in light of the framework (for other arguments against the internalist reading of PP, see Clark, 2016a, 2017; Fabry, 2017). However, two things need to be pointed out before the conservatist admits defeat on this front. First, the internalist commitment is logically independent from the other two. Most importantly, neither representationalism nor inferentialism about PP presuppose the truth of internalism. There is nothing contradictory about the idea of a system that trades in representations and engages in inferences but whose boundaries do not coincide with the boundaries of the central nervous system. So even if we do drop the internalist commitment, the other two can remain intact, leaving us with what is still a recognizably (albeit weakly) conservative outlook on the nature of cognition. Second, even if internalism cannot be defended by pointing to the notion of a Markov blanket alone, there may be other, independent considerations in favor of internalism. In particular, it might be interesting to see how PP meshes with other, independent theoretical proposals that support delineating cognition in internalist, skull-bound way. A full, in-depth discussion of this subject is beyond the scope of this paper. However, let me briefly sketch out the connections beyond PP and some of the well-known, pro-internalist conceptions of where cognition ends.

i. PP and non-derived content

On one view, what distinguishes cognition from non-cognition is the fact that only the former involves processes that make use of *non-derived intentional content* (Adams & Aizawa, 2001, 2010). This is the content that

is intrinsic to the content-bearing state rather than derived from conventions or interpretative/explanatory practices. Note that when applied to PP, this approach would connect internalist commitment to the representational one. Because on the weak, pragmatist/instrumentalist reading of representationalism in PP, content is clearly *derived* (it depends in its existence on the explanatory practices of scientists), the connection would have to be with the *strong* branch of representationalism. The internal, resemblance-based, action-driving model that the strong reading of representations in PP appeals to seems like a good seat for non-derived content. The content of this model is based on the structural resemblance between the representational vehicle and some (represented) part of the environment, such that the degree to which the resemblance holds is causally relevant for the success of model-guided actions (Gładziejewski, 2016; Kiefer & Hohwy, 2017; Williams, 2017). Neither the structure of the vehicle, the structure of the represented state of affairs nor the resemblance relation itself are observer-dependent; this view of content is realist through and through (see also Gładziejewski & Miłkowski, 2017). Hence, it is reasonable to assume that content here is not of derived nature. Assuming further that the generative model that serves as representation turns out properly situated within the confines of the skull, we end up with an internalist view. The weakness of this proposal lies in the non-derived-content-based strategy of delineating cognition itself. By definitionally linking cognition with representational content, this criterion is hardly ecumenical towards 4E approaches. More importantly, it seems to deflate or trivialize representationalism by *a priori* precluding the truth of anti-representationalism about cognition (Ramsey, 2015).

ii. PP and cognitive systems

Another internalist way of demarcating cognition appeals to the notion of a cognitive system (Rupert, 2009). Roughly, 'cognitive systems' are physical systems that causally underlie collections of cognitive capacities and skills. These systems are integrated and persisting, and the collections of cognitive capacities and skills they give rise to are stable across different contexts. Because of their persisting and stable nature, it is cognitive systems that enable successful psychological or cognitive-scientific explanation by

making possible reliable generalizations about cognition. They give rise to stable patterns of cognitive behavior that can be studied under a wide range of independent experimental paradigms. The proposal is that only brains (or central nervous systems) count as 'cognitive systems' in this sense. For example, it is argued that 'extended' systems which comprise the (embodied) brain and parts of the environment are too ephemeral to afford successful, generalizable scientific inquiry (Rupert, 2009). Now, it might be hypothesized that the central nervous system *qua* prediction-error-minimizing mechanism counts as a cognitive system in this sense. It persists across different contexts and gives rise to cognitive phenomena. Furthermore, it might be argued that although there are extended prediction-error-minimizing systems enclosed by technology-based Markov blankets, these are not cognitive, as they are not stable enough to underlie successful scientific generalization. If this is true, it could rule out Clark's metamorphically extended predictive minds. The obvious problem, however, is that there are Markov blanketed, prediction-error-minimizing mechanisms *within* the central nervous systems. These may be even more stable and persisting error minimizing mechanisms within the agent. So, there remains something arbitrary about treating the peripheries of the central nervous system as the peripheries of cognition.

iii. PP and pseudo-closed-loop control

Grush (2003) defends internalist or 'Cartesian' demarcation of cognition by employing notions from control theory. To put Grush's sophisticated account in a nutshell, the idea is that brains count as sole seats of cognition because they are systems for which 'the world is not enough'. Due to the temporal delay that separates the sending of a motor command to the body and the sensory feedback resulting from the performed action, the brain is unable to perform motor control based on the feedback alone. Rather, it uses a pseudo-closed-loop architecture, where an efference copy of motor command is sent to an emulator, an internal structure that mimics the dynamics of the environment and the muscle-skeletal system. The sensory predictions endogenously derived from the emulator are essential, on Grush's account, for planning and fine-tuning ongoing movement. Furthermore, the emulator can be employed for purely off-line purposes,

like imagery. The upshot is that because of its reliance on an internal emulator, the brain emerges as 'potentially self-contained' – a system firmly distinct from the external environment and (given some other assumptions, see Grush 2003) a unique seat of cognition. Now, there is a recognizable kinship between emulation theory (and other efference-copy-based approaches to motor control) and PP (Dolega, 2017). Most importantly, note how perception and action are crucially guided in PP by endogenously generated, top-down sensory predictions. Obviously, because of the crucial role that the sensory input and error signals have in shaping the internal processing, the prediction-error-minizing system is far from being closed-off from the environment. This does not, however, diverge from Grush's original emulation framework, as, on his view, the sensory feedback constantly corrects the emulation-based predictions. Notice also how, in PP, when the precision of the sensory signal is predicted to be low (and so the sensory input's influence on hypothesis-revision is also low), or when the generative model is used purely off-line, the brain will appear as largely causally decoupled from the external environment. Because of those considerations, there is potential in PP to construe the brain (or the central nervous system) in Grushian way, as a largely self-contained seat of cognitive phenomena.

Conclusions

When seen within the proper context of the Free Energy Principle, minimizing the prediction error with the use of hierarchically structured generative models turns out to serve as a tool for self-organization. This strong pragmatic and organism-oriented spin on PP naturally invites an interpretation of the framework that is much closer to 4E approaches than to more orthodox, internalistic and intellectualist approaches in cognitive science. However, in the present paper I attempted to elucidate what 'conservative' reading of PP amounts to, hoping to show that this way of understanding PP is not ungrounded or completely alien to the spirit of the 4E approaches. I showed how PP is representational, both in a weak (pragmatic) and strong (realist) sense. Even on the strong reading, the representations postulated in PP are not just passive mirrors of nature, but action-guiding map-like structural representations that largely use

modality-specific vehicles and whose content is constrained by the way the organism is embodied and embedded in its environmental niche. Furthermore, I argued that the notion of inference that (the conservative rendering of) PP trades in is non-trivial, yet liberal. The inferential nature of perception amounts to the fact that the way the perceptual representations are (actively, not just reactively) updated conforms to a truth-preserving rule. There is no commitment here to an overly intellectualist claim that prediction-error-minimizing agents cognize in accordance to some inflated principle of rationality. Lastly, I attempted to show that whatever grounds there might be for treating PP in internalist terms, they are probably not to be found in the conceptual resources of the framework itself. However, I sketched out how PP might fit with some other, independent ways of delineating the mind in a skull-bound way. The resulting view is that PP is representational and inferential in what might be the most 4E-friendly way possible, and it does not have to be considered internalist (at least not on its own terms). Taken together, these considerations show that conservative reading of PP is well-grounded and not *that* conservative after all.

REFERENCES

Adams, F., Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, 14, 43–64.

Adams, F., Aizawa, K. (2010). *The Bounds of Cognition*. Oxford: Wiley-Blackwell.

Allen, M., Friston, K. J. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, doi: 10.1007/s11229-016-1288-5.

Anderson, M. L. (2017). Of Bayes and bullets. In T. Metzinger, W. Wiese. *Philosophy and Predictive Processing*. MIND Group, ISBN: 9783958573055.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-609.

Bickhard, M. H. (1999). Interaction and Representation. *Theory & Psychology*, 9, 435–458.

Bruineberg, J., Kiverstein, J., Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, doi: 10.1007/s11229-016-1239-1.

Burr, C., Jones, M. (2016). The body as laboratory: Prediction-error minimization, embodiment, and representation. *Philosophical Psychology*, 29, 586–600.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 181–204.

Clark, A. (2015). Radical predictive processing. *The Southern Journal of Philosophy*, 53, 3–27.

Clark, A. (2016a). Busting out: predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Nous*, doi: 10.1111/nous.12140.

Clark, A. (2016b). *Surfing Uncertainty. Prediction, Action, and the Embodied Mind.* Oxford: Oxford University Press.

Clark, A. (2017). How to knit your own Markov blanket : Resisting the second law with metamorphic minds. In: T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*, MIND Group, ISBN: 9783958573031.

Dolega, K. (2017). Moderate predictive processing. In T. Metzinger, W. Wiese (eds), *Philosophy and Predictive Processing*. MIND Group, ISBN: 9783958573116.

Downey, A. (2017). Predictive processing and the representation wars: a victory for the eliminativist (via fictionalism). *Synthese*, doi: 10.1007/s11229-017-1442-8.

Egan, F. (2010). Computational models : a modest role for content. *Studies in History and Philosophy of Science*, 41, 253–259.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, 170, 115–135.

Fabry, R. E. (2017). Transcending the evidentiary boundary: Prediction error minimization, embodied interaction, and explanatory pluralism. *Philosophical Psychology*, 30, 391–410.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews. Neuroscience*. 11(2), 127–138.

Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, 10, 20130475–20130475.

Friston, K. J., Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, 159, 417–458.

Gładziejewski, P. (2015). Explaining cognitive phenomena with internal representations: A mechanistic perspective. *Studies in Logic, Grammar and Rhetoric*, 40, 63–90.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, 193, 559–582.

Gładziejewski, P. (2017). The evidence of the senses: A Predictive Processing-based take on the Sellarsian dilemma. In T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*, ISBN: 9783958573161.

Gładziejewski, P., Miłkowski, M. (2017). Structural representations: causally relevant and different from detectors. *Biology and Philosophy*, 32, 337–355.

Goldman, A. I. (2012). A moderate approach to embodied cognitive science. *Review of Philosophy and Psychology*, 3, 71–88.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 290, 181–97.

Grush, R. (2003). In defense of some 'Cartesian' assumptions concerning the brain and its operation. *Biology and Philosophy*, 18, 53–93.

Helmholtz, H. (1860/1962). *Handbuch der Physiologischen Optik*. J P. C. Southall (ed), Vol. 3. New York: Dover.

Hobson, J. A., Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, 98, 82–98.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Hohwy, J. (2016). The self-evidencing brain. *Nous*, 50, 259–285.

Hohwy, J. (2017). How to entrain your evil demon. In T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*, MIND Group, ISBN: 9783958573048.

Hohwy, J. (in print). The predictive processing hypothesis and 4e cognition. In A. Newen, L. Bruin, S. Gallagher (eds), *The Oxford Handbook of Cognition: Embodied, Embedded, Enactive and Extended.*

Hohwy, J., Roepstorff, A., Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108, 687–701.

Hutto, D. D. (2017). Getting into predictive processing's great guessing game: Bootstrap heaven or hell? *Synthese*, doi: 10.1007/s11229-017-1385-0.

Kiefer, A. (2017). Literal perceptual inference. In T. Metzinger, W. Wiese (eds) *Philosophy and Predictive Processing*, ISBN: 9783958573185.

Kiefer, A., Hohwy, J. (2017). Content and misrepresentation in hierarchical generative models. *Synthese*, doi: 10.1007/s11229-017-1435-7.

Klein, C. (2016). What do predictive coders want? *Synthese*, doi: 10.1007/s11229-016-1250-6.

Orlandi, N. (2016). Bayesian perception is ecological perception. *Philosophical Topics*, 44, 255–278.

Orlandi, N. (2017). Predictive perceptual systems. *Synthese*, doi: 10.1007/s11229-017-1373-4.

Ramsey, W. (2015). Must cognition be representational? *Synthese*, doi: 10.1007/s11229-014-0644-6.

Pezzulo, G. (2017). Tracing the roots of cognition in predictive processing. In T. Metzinger, W. Wiese (eds), *Philosophy and Predictive*

Processing, MIND Group, ISBN: 9783958573215.

Rescorla, M. (2013). Bayesian perceptual psychology. In M. Matthen (ed.), *The Oxford Handbook of Philosophy of Perception* (694–716). Oxford University Press.

Rosenberg, D. G., Anderson, M. L. (2004). Content and action: The guidance theory of representation. *The Journal of Mind and Behaviour*, 29, 55–86.

Rupert, R. (2009). *Cognitive Systems and the Extended Mind*. Oxford: Oxford University Press.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17, 565–573.

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, 5, 97–118.

Sprevak, M. (2013). Fictionalism about neural representations. *The Monist*, 96, 539–560.

Wiese, W. (2016). Action is enabled by systematic misrepresentations. *Erkenntnis*, doi: 10.1007/s10670-016-9867-x.

Wiese, W. (2017). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, 16, 715–736.

Wiese, W., Metzinger, T. (2017). Vanilla predictive processing for philosophers: A primer on predictive processing. In T. Metzinger, W. Wiese (eds), *Philosophy and Predictive Processing*. MIND Group, ISBN: 9783958573024.

Williams, D. (2017). Predictive processing and the representation wars. *Minds and Machines*, doi: 10.1007/s11023-017-9441-6.

ABSTRACT

JUST HOW CONSERVATIVE IS CONSERVATIVE PREDICTIVE PROCESSING?

Predictive Processing (PP) framework construes perception and action (and perhaps other cognitive phenomena) as a matter of minimizing prediction error, i.e. the mismatch between the sensory input and sensory predictions generated by a hierarchically organized statistical model. There is a question of how PP fits into the debate between traditional, neurocentric and representation-heavy approaches in cognitive science and those approaches that see cognition as embodied, environmentally embedded, extended and (largely) representation-free. In the present paper, I aim to investigate and clarify the cognitivist or 'conservative' reading of PP. I argue that the conservative commitments of PP can be divided into three distinct categories: (1) representationalism, (2) inferentialism, and (3) internalism. I show how these commitments and their relations should be understood and argue for an interpretation of each that is both non-trivial and largely ecumenical towards the 4E literature. Conservative PP is as progressive as conservatism gets.

KEYWORDS: embodied cognition; enactivism; Free Energy Principle; inference; internalism; Predictive Processing; mental representation

HYBRIS nr 38 (2017)
ISSN: 1689-4286

KRYSTYNA BIELECKA
UNIVERSITY OF WARSAW

SEMANTIC INTERNALISM IS A MISTAKE

In this paper, I introduce the concept of *narrow content* (Section 2.1) to discuss an account of narrow content by analyzing Fodor's methodological solipsism (2.2). I point out that Fodor's formalism, that is, the position according to which the content is reduced to formal properties of mental representation, eliminates (at least - as I show in Section 2.2.4 - in Stich's interpretation) semantic properties in favor of the syntactic ones. In addition, it leads to the conceptual problems indicated by J. Searle, S. Harnad (Section 2.3), and T. Burge (Section 2.4). In a nutshell, semantic internalism, as reviewed in this paper, does not offer an account of content that would be properly contentful, because it provides no grounds to ascribe truth or other semantic properties to representations. In particular, it is either unsatisfactory, because it reduces content to formal properties or inconsistent, because it appeals to innate contents that itself has not been properly explicated; moreover, innate factors, as I argue, are not merely individual. Consequently, I reject semantic internalism in favor of externalism.

The purpose of this paper is to argue against the usefulness of narrow content in the account of mental representation. By reviewing the classical arguments in favor of the narrow content, I show that the notion is inevitably wrong-headed. This is probably the reason why even one of the most radical proponents of narrow content, J. Fodor, changed his mind and rejected the narrow content in favor of wide content (Fodor 2008). Any future effort of defending the notion of narrow content will have to face the challenge of demonstrating that the narrow content has semantic properties.

2.1. The philosophical notion of intension and extension

It is generally assumed that there is an analogy between linguistic meaning and content of mental representation (Pitt 2013). Traditionally, intension or connotation (meaning for linguistic expressions, content for mental representation) and extension (mental representation can be about something, true or false about an object, or true or false *simpliciter*) are attributed to mental representations and linguistic expressions.[1] The correspondence between language and thought with reality can justify this analogy; both linguistic expressions and mental representations refer to reality and describe it.

A complex expression is extensional if its denotation is a function of denotations of its constituent expressions (Jadacki 2001). In extensional contexts, substituting one of the constituents of an extensional expression with a constituent with the same denotation does not change the logical (or semantical) value of the whole expression. Knowing the denotation of constituents of an extensional expression is sufficient to determine the logical (or semantical) value of the whole expression. I'll illustrate this by two sentences:

a) George W. Bush voted for Barack Obama in 2012.

b) The last but one ex-president of the US voted for Barack Obama in 2012.

Sentence b) was formed as a result of substituting one of the constituents of an extensional expression, "George W. Bush", with a constituent with the same denotation, "the last but one ex-president of

[1] Connotation is the minimal set of properties related to a language expression X that anyone speaking the language to which X belongs can use to recognize the referent of X. The notion of connotation is traditionally used interchangeably with a concept of intension (Copi and Cohen 2002). However, these concepts are sometimes differentiated. This happens when one assumes a slightly different sense of intension. In the Carnapian tradition, intension is the function of language expressions onto noncontradictory sets of propositions (Carnap 1947), and in tradition of two-dimensional semantics it is a function onto possible worlds (Chalmers 2004). This function assigns the extension to a term (in a given possible world). For example, in our possible world, the terms "Evening Star" and "Morning Star" have the same intension across contexts, but different connotations.

the US", but the logical value of the sentence did not change. Both expressions refer to a specific person, that is the former US president, George W. Bush.

An expression is intensional (referentially opaque) if and only if it is not extensional. In expressions occurring in intensional contexts, substitution of a selected constituent of a sentence with another constituent of the same logical (or semantical) value may change the logical value of the whole sentence. Intensional expressions include such sentences as "x thinks that p", "x knows that p", "x wants p", "x believes that p". Suppose that George W. Bush voted for Barack Obama, but Johnny does not know that George W. Bush is the last but one ex-president of the US. I'll illustrate this by the following example:

a) George W. Bush voted for Barack Obama in 2012.

b) The last but one ex-president of the US voted for Barack Obama in 2012.

Both constituent parts, namely "George W. Bush voted for Barack Obama in 2012" (in the sentence a*) and "The last US expedition voted for Barack Obama in 2012" (in the sentence b*), have the same logical value, but sentence a* may have a different logical value than b* if Johnny thinks that the last but one ex-president of the US and George W. Bush are two different people.

2.2. Naturalized conceptual role semantics

A naturalistic account of content of mental representation that explicates the content in terms of their functional roles in a cognitive system is a promising attempt to provide semantics for psychology. By a functional role, it is generally meant:

> the role of that representation in the cognitive life of the agent, e.g. in perception, thought and decision-making (Block, 1998).

This definition of a conceptual role is, however, very general. It can be accepted both by internalists, who adopt the notion of narrow content, as well as externalists, who embrace wide content. Narrow content is limited to the functional role within the cognitive system, while wide content also includes the context and environmental circumstances in which the cognitive system is situated.

2.2.1. In search of narrow content. Cognitive nature of content

According to internalists, the concept of narrow content suffices to describe, explain, and predict the intentional states of all cognitive systems. They consider the concept of wide content to be defective and useless in psychology, as the concept seems to be too dependent on the circumstances of its occurrence and insufficiently dependent on the structure of the cognitive system itself. In order to evaluate these theses, I will closely examine the concept of narrow content, the arguments evoked for its defense, and intuitions underlying internalist views.

Internalists point out to the mind's ability to think about things that are not the case. I will illustrate such intuitions by an example of Johnny standing at the bus stop waiting for the bus. If he has an incorrect timetable, he may still think of the bus, even though a vehicle is not arriving and will not arrive. Moreover, the same thought about the bus can reappear in the boy's head at various times and at different places: both at the bus stop and at home. So, it would seem that thoughts preserve their content regardless of their context (place or time). Johnny can also think at any time of a mountain of gold, even though he knows perfectly well that it does not exist. Thoughts also retain their content when they do not refer to anything real.

The concept of narrow content preserves the independence of content of thought with regard to such factors as reference and satisfaction conditions. Thus, an internalist G. Segal describes it as *cognitive* content in contrast to referential/truth-conditional content (Segal 2000).

Internalism opposed externalism, including Wisconsin style informational-causal theory of reference, which rejects the autonomy of content from reference. The idea of extreme externalism was spelled out by H. Putnam in his Twin-Earth thought experiment (Putnam 1975). Putnam shows that intension does not determine extension. Imagine that on a distant Twin Earth, a liquid called "water" does not refer to a chemical substance H_2O, but to XYZ. Water from Earth is indistinguishable from water on Twin Earth: it looks the same and tastes the same. Imagine that before the discovery of the chemical structure of water, before 1750, on Earth, there was $Oscar_1$, who had a twin brother on the Twin Earth, $Oscar_2$. Oscars did not distinguish XYZ from H_2O and for both the intension of the expression "water" was the same: a colorless, potable liquid. Putnam claims that the meaning of the

word "water" as used by $Oscar_1$ was not the same as the meaning of a word "water" uttered on the Twin Earth, because these expressions differed in their denotation. It is not the case that intension fully determines extension, because the environment is also critical for meaning.

Internalists claim that there was a common meaning of the word "water" on both planets, which is not reflected in claims of externalists, according to which the content of the same term was different on both planets. Their claim overlooks, according to internalists, an essential aspect of mental content thanks to which one can predicate the same attributes about the same object independently of external factors.

2.2.2. Determination of narrow content in terms of supervenience

According to Block's definition, internalists propose to characterize content only in terms of its causal-inferential roles within an individual cognitive system. Narrow content is therefore a part of the internal structure of an individual cognitive system; it participates in its cognitive life – in its inferences, decision-making, and so on (Block 1987). The fact that it is all about the internal structure of an individual cognitive system is emphasized by Segal (who defines narrow content in terms of local supervenience), according to which narrow content is a property of mental representation completely independent of any external factors. Internalism assumes that narrow content can only be determined by the internal structure of an individual cognitive system.

The content of mental representation is entirely determined by intrinsic properties of an agent or a cognitive system. An intrinsic property is a property that the object has (or not) regardless of what is the case beyond that object (Yablo 1999). An example of an intrinsic property is a square's property of having-four-equal-sides: a square always has four equal sides, regardless of how things are outside it. On the other hand, being a living organism is not an intrinsic property of an organism, because the organism would not have this property if it didn't, for example, breathe oxygen. And so, intrinsic properties of a cognitive system are those properties of the internal structure of the cognitive system that remain independent of any external factors of the system. To talk of such properties, I will use a term *microstructural properties*. Microstructural properties are properties of an internal structure of a cognitive system and its parts (and relationships between

them) - unlike macrostructural properties which include relationships with the environment and other agents.[2]

The claim about the determination of narrow content expressed in terms of local supervenience states that contentful properties depend only asymmetrically on microstructural properties of a cognitive system. Segal characterizes narrow content in terms of local supervenience on the microstructure of a cognitive system. He argues that microstructural properties are sufficient to determine the neural and computational properties of a system, i.e., narrow content:

> Fix an object's microstructure and you fix its atomic and molecular structure, its neurological and computational properties, and so on. (Segal 2000, 14).

The definition of narrow content in terms of local supervenience allows us to explicate more precisely how content is determined according to internalism, and at the same time, to indicate a problem related to the determination understood this way. Local supervenience does not allow us to precisely define the character of narrow content, since it does not provide a way to define properties independent from the external environment to the system, and narrow content depends only on these properties. On the contrary, it excludes only certain groups of (externalistic) accounts. Moreover, the consequence of a local supervenience claim is that all cognitive interactions with environment are irrelevant for content, which is fully reducible to the microstructure of the system.

Here, an ontological reduction is at stake, that is, a relation between elements of the real world, such as objects, events or properties. It occurs if relationships, such as elimination, identity, superposition, realization, or supervenience occur (van Gulick 2001). Because in internalism supervenience between properties is assumed, it is an example of such an ontological reduction.

---

[2] The terms *microstructural* and *macrostructural* have been proposed by R. Poczobut, who formulated the supervenience claim in terms of micro and macrostructure (Poczobut 2007).

2.3. Fodorian methodological solipsism

I will now concentrate on Fodor's defense of the concept of narrow content in his methodological solipsism (Fodor 1980; Stich 1980). Narrow content in Fodor's view arises from the reduction of semantic properties to syntactic ones and to innate semantic properties as well. This reduction, is not, however, a full naturalization, as it is unclear how innate content is determined. Thus, only a total reduction of the content to syntactic properties is fully consistent and naturalistic, but such an account on narrow content deprives it, alas, of its content. Consequently, the syntactic understanding of content is - contrary to what Fodor claims - inadequate for psychology and cognitive sciences. Thus, conceptual role semantics that would accept only narrowly understood content would not be a semantics for psychology.

2.3.1. Narrow content in methodological solipsism

Fodorian methodological solipsism plays a key role in developing the concept of narrow content in psychology and philosophy of psychology. Within this framework, Fodor attempts to defend methodologically individualist psychology as the only proper approach to psychological research. He responds to Putnam's counterarguments against the classical claim that intension determines extension [3]

Narrow content in methodological solipsism is characterized by inferential roles that are syntactic and computational.[4] Fodor argues that knowledge of intrinsic properties, especially formal representational properties suffices to describe the content of a

[3] Later Fodor did not link individualism in psychology and methodological solipsism so closely. Moreover, he distinguished between methodological solipsism and methodological individualism (Fodor 1987). According to methodological solipsism, mental states are individuated without semantic valuation; an (externally) relational taxonomy of mental states is methodologically unacceptable. Methodological individualism, on the contrary, allows for relational individuation of mental states provided that a mental state property is only included in the mental states taxonomy if it is causally relevant (Heath 2015).

[4] Identification of inferential roles with causal ones stems from the classic computational account of functionalism. Inferential roles characterized syntactically are roles in a computational architecture of mind (Field 1978; Fodor 1975).

representation. He claims that it suffices to express all content relevant to cognitive psychology.

According to Fodor, a representational relation has two related members. It consists of a relation to a sentential object, described formally (syntactically), and of a relation to this object, described semantically in terms of truth and reference. However, as Fodor adduces, formal properties of narrow content are sufficient to distinguish semantically different representations. Fodor is, therefore, a proponent of a formalist account of a theory of content, whereby formal properties of signs (e.g., their shapes or structure) and syntactic rules are sufficient to characterize content.

Fodor describes the mind as a so-called oracle machine, where "oracle" is understood, after Turing, as a procedure that settles a question in a non-computational (non-algorithmic) manner. According to Fodor, the role of oracle is played by the perceptual states of an environment:

> The point is that, so long as we are thinking of mental processes as purely computational, the bearing of environmental information upon such processes is exhausted by the formal character of whatever the oracles write on the tape. In particular, it doesn't matter to such processes whether what the oracles write is true; whether, for example, they really are transducers faithfully mirroring the state of the environment, or maybe the output end of a typewriter manipulated by a Cartesian demon bent on deceiving a machine (Fodor 1980, 65).

Fodor claims that a formal difference makes a functional difference, which in turn makes a causal difference:

> The form of explanation goes: it's because different content implies formally different internal representations (via the formality condition) and formally distinct mental representations can be functionally different; can differ in their causal role. Whereas, to put it mildly, it is hard to see how internal representations could differ in causal role unless they differ in form (Fodor 1980, 68).

For Fodor, this is a pragmatic argument for substituting formal properties with semantic ones in explanations.

Fodor, like Davidson, defends folk psychology, and therefore points out that statements about beliefs, thus referentially opaque

contexts, are crucial in folk psychology. He even claims that folk psychology does not need anything more than an explanation of representation in opaque contexts, in which generalizations are about what people mean by propositions to which they express propositional attitudes.

Propositions that occur in opaque contexts differ not only in content but also in their form. That is why such sentences do not undermine the Fodor's assumption that only intrinsic structural properties have causal powers. Formal properties suffice, according to Fodor, to adequately describe the content, such as beliefs, even in opaque contexts. Later in this paper, I will argue against Fodorian concept of narrow content, according to which a formal difference is sufficient to explain the difference in intension.

2.3.3 Concept innateness

Fodor defends his own account of concepts, which is a philosophical interpretation of a classical theory of concepts, enriched by an innateness hypothesis. In a classical theory, the concept is:

> a class representation, covering all relevant properties of such class. A criterion of relevance is repetition - an essential feature is the feature that characterizes all objects belonging to this class, i.e., - in other words - the common feature of all objects belonging to this class (Maruszewski 1983).

Maruszewski's definition differs just a little from the classical definition of connotation, according to which connotation of *z* is a property which applies to all *z*-s and only *z*-s (see also Jadacki 2001, 107).

The innateness hypothesis states that our basic conceptual apparatus is innate.[5] We have a language of thought, that is, an internal

[5] The concept of innateness is unclear and can be understood in many different ways, especially since in contemporary psychology one does not disregard the biological basis of cognition. In biology, it is by no means clear not only what is innate but also what innateness is; philosophers of biology challenge the utility of such a concept (Samuels 2002, 2004; Griffiths 2002). One of the explications of the concept of innateness on the biological ground defines innateness as a disposition to behave under normal conditions. This explication, however, must go beyond narrow content, as normal conditions always appear in an environment.

code, to which all other concepts can be reduced, and we can perform proper combinatorial operations (Fodor 1975). For example, a non-basic concept of BORING BOOK consists of basic concepts BOOK and BORING. One of Fodor's most important arguments for the language of thought hypothesis in psychology is the productivity of thought. Language is one such productive system. Fodor has to justify why it is language and not another productive representation system that is innate. The key argument for language innateness serves this purpose.

Fodor (1975), analyzing psychological theories of concept learning, notes that they all regard learning concepts as a process of hypothesizing. Hypotheses can only be posed in language, and, therefore, in order for a child to pose them she must have an innate language of thought. Before she can learn concepts she must be able to formulate such hypotheses. That is why language, and not a different productive representation system, is innate: in a different system such hypotheses cannot be made.

Let's go back to productivity. It is obvious that we can also think an infinite number of different thoughts; similarly, we can utter an infinite number of sentences, for example "Giraffes do not play poker". According to Fodor, it is impossible to explain the productivity of language and thought without assuming the compositionality of language. It is language that is so rich in structure that makes it productive. Therefore, as he claims, thinking occurs in a linguistic medium.

A special case of concepts in the Fodorian account are concepts that can no longer be broken into constituent parts. These include concepts for simple sensory qualities, i.e., shapes and colors, and the simplest colloquial concepts. The structure of a concept also plays a role in deciding whether a concept belongs to a basic category: it must be a concept without which other concepts cannot be created in virtue of a compositional principle. For example, a concept BORING BOOK is a non-basic complex concept consisting of a basic concept BOOK and a basic concept BORING - concepts BOOK and BORING cannot be simplified further and these are concepts without which creating a concept BORING BOOK would be difficult.

By design, methodological solipsism forbids citing environmental properties in explaining the determination of the content of such basic concepts. Thus, they remain unexplained, and the

Fodorian account can never really explain how their content was determined. The content of basic concepts remains an aporia of methodological solipsism. Within methodological solipsism, determining the content of basic concepts is logically impossible: if they are innate, they depend not only on the individual cognitive agent but also on biological inheritance, which goes beyond the agent.

The hypothesis of an innate conceptual apparatus could remain consistent with the local supervenience of such narrow content on the formal properties of a conceptual apparatus, if only the formal properties of this apparatus constituted the basis of conceptual content. Fodor, however, is opposed to the total reduction of semantics to formal properties. One way to solve the problem of the determination of content of basic concepts, which would allow him to preserve methodological solipsism, is to treat it precisely as a formalistic account, that is, purely syntactic. As a result, this could strengthen and radicalize Fodor's account.

Fodor does not accept the strong claim that all semantic properties can be reduced to syntactic ones, but he claims something weaker: formal properties are the most satisfactory indicator of content. I do not intend to argue with that: indeed, the form is the simplest indicator of difference in content and, in addition to this, it works well in many situations.

2.3.4. Methodological solipsism and a formalist account

In this section, I will examine Stich's more radical account which argues for a complete reduction of semantics to syntax and, consequently, for the elimination of the notion of representation from folk psychology. Stich indicates to what exactly the reduction of content to formal property leads. In essence, Stich shows that an internalist concept of content is not a concept of content. This means that internalism cannot naturalize intentionality.

Stephen Stich starts from a formalist interpretation of methodological solipsism, criticizing Fodor's argument for this position. Stich's counterargument can be understood as being directed against methodological solipsism as well as against psychology that uses the notion of representation. Stich's objection to Fodor's notion of narrow content shows that this notion is divergent from its folk counterpart and, consequently, cannot be used to defend folk psychology. Stich thus

argues for rejecting any concept of content. In addition, he is against representationalism because he thinks - as Fodor does - that computational psychology does not adhere to the principle of charity and thus, does not take the semantic properties of representations, such as truth, into account.

Stich argues against Fodor that his concept of narrow content leads to undesirable consequences:

> First, most computational (or formal) mental states will have tokens (either actual or possible) whose contents are radically different from one another, as judged by our "aboriginal, uncorrupted, pretheoretic intuition". Second, there will be some computational mental state types whose tokens can be assigned no content at all by our aboriginal intuitions, though these "contentless" computational states will serve the purposes of the computational theory of mind fully as well as their contentful cousins (Stich 1980, 97).

First, Stich observes the concept of narrow content leads to ascribing the same content in intuitively different cases. Second, he stresses that an account of narrow content typical for methodological solipsism leads to one more undesirable consequence: it does not forbid attributting content to beliefs radically diverging from our own beliefs, even if they violate common intuitions about their content.

The cases of the first kind are analogous to Putnam's example of Twin Earth. I will mention one of them. In Stich example, Fodor from Yon (Putnam's Twin Earth analogue) appears and utters, analogically to Fodor from Earth: "Jimmy Carter is from Georgia." On the Twin Earth far apart from our Earth, even further than Yon, there could be a Twin Fodor, uttering the same sentence "Jimmy Carter is from Georgia". But then, according to the Fodorian account of narrow content, the same content should be attributed to this sentence on Twin Earth. Stich notices that this is completely inconsistent with a common intuition that different Fodors, depending on where they are located, speak of a different Jimmy Carter, depending on where they come from (assuming that on Yon and on Twin Earth there is also Georgia). I agree with Stich that the context of utterance should be taken into account while assigning content to a belief. Indeed, defending an account of reference of proper names requires one to accept a causal account of content that is incompatible with methodological solipsism. Dependence of content

on context is one of the reasons for adopting the wide account of content.

The notion of narrow content should also be attributed to such things or creatures to which the principle of charity would forbid attributing any content. Stich illustrates this with examples of robots whose beliefs are so distant from ours that one can never agree that, according to the principle of charity, their representational states have any content. If there was a robot simulating a human being with beliefs vastly contradicting one other, the principle of charity would not allow us to attribute content to it, because, in such a case, neither truth nor accuracy of its "beliefs" could be treated seriously. Fodor has no way to deny that these "beliefs" are meaningful by his own lights.

In conclusion, Stich shows that a formalist account does not lead to a defense of folk psychology but rather to its rejection. Although he agrees with Fodor that a formalist account suffices to describe content, he goes a step further, claiming that a notion of representation should be rejected totally from folk psychology. At the same time, according to him, we must reject a notion of narrow content and substitute it with a notion of form. I agree with Stich's argument against a notion of narrow content if narrow content is reduced to formal properties. Such a reduction does not properly describe content in contexts in which expressions differ in content but not in form. I propose, however, to treat Stich's argument - contrary to his intentions - as a warning against an excessively hasty reduction of content to form and against the elimination of the concept of mental representation.

Narrow content in methodological solipsism will not allow us to distinguish between representations whose form does not decide their meaning. This group includes homonyms (such as "bank") and representations whose meaning depends on the environment (Putnam's example of water on Twin Earth). Their form is the same, but the content is different because: (A) both intension and extension are different (in the case of homonyms); or (b) extension (of representations whose content depends on the environment) is different. A formalist account could deal with homonyms at the level of expressions, denying *de facto* their existence: by separating those sentences where the word "bank" in the meaning of "a bench of the river" is different from sentences in which there is a "bank" in the meaning of "a building in which you put your money" because of

inferential roles that these words play in sentences, and the roles determine the meaning of the word "bank". For formalists more difficult are homonyms at the level of the sentence, where a pragmatic context plays a decisive role: the sentence "You have huge feet!" uttered in a shoe store expresses the fact that someone has feet of large size, but uttered in the presence of someone with small feet is ironic. One should also remember the role played by the intonation: the same sentence said in a dismissive tone can be offensive (Dennett 1991). In the case of homonyms at the level of the whole sentence, resorting to inferential roles is practically infeasible - it is impossible to distinguish sentences that are so strongly contextually dependent, purely on a syntactic level, thus invoking only their formal properties.

The formalist account may, however, show the difference in meaning between the sentences "George W. Bush voted for Barack Obama in 2012" and "The last but one ex-president of the US voted for Barack Obama in 2012," because the term "George W. Bush" has a different form than the term “the last but one ex-president of the US".

The formalist account is valid in many interesting cases, in which the form of a vehicle corresponds to representational content. It would be a mistake to ignore formal properties in an account of content. However, in order for a formal account to fully replace semantics, it should be able to explain the cumbersome cases described above.

Thus, the adequate account of content should not be a purely formal conception of content, since such an account is powerless in those contexts, in which the reference clearly decides about content. Twin Oscar's statement about water is therefore considered to be different from Earthly Oscar's statement in which "water" refers to a different chemical structure on Earth than the one to which it would refer on Twin Earth. It is therefore reasonable to assume that an adequate account of content should not be merely formal.

2.4. Arguments against the formalist account

2.4.1. Chinese Room thought experiment

The most well-known polemic against supporters of a formal account of representation, in this case symbolic representation, can be found in J. Searle's Chinese Room thought experiment (Searle 1980). Searle who doesn't know any Chinese is enclosed in a room with a text file in

Chinese ("script") with some additional files: a set of rules (equivalent to a program) correlating the second file with the first one (called by Searle "a story") and a set of instructions given in English (questions), allowing to correlate elements from the third file with the first two. These instructions specify how to send certain Chinese symbols of certain shapes, responding to those shapes in a third file. Both the program and answers to such questions are for Searle purely syntactic transformations of symbols. Searle, in his room, is supposed to answer the questions given sometimes in English, and sometimes in Chinese. As it turns out, his answers to the questions in Chinese do not differ from those of a Chinese man who natively speaks Chinese. Additionally, the answers to questions posed in Chinese are as good when seen from the outside as the answers to questions in English. In the first case, Searle's answers are only non-interpreted symbols.

Searle's thought experiment is supposed to deal with many philosophical issues, including consciousness, artificial intelligence, and mental representation. I will focus on the last issue, and within it, on the subject of my interest: the possibility to characterize content solely formally. For this purpose, I will appeal to one of many formulations of Searle's argument, and within it, to the interdependence of form and content (Hauser 1997). The argument has the following form:

1. Programs are purely formal.
2. Minds (or at least human minds) have semantics, mental content.
3. The syntax itself does not constitute content nor is a sufficient condition for content.

Programs themselves are neither constitutive nor sufficient for the functioning of mind. (Preston and Bishop 2002).

Many philosophers question the validity of Searle's argument (Dennett 1987; Chalmers 1996). I think, however, that the core of Searle's argument, that is the claim that syntax is neither identical to content, nor sufficient to describe semantic content, points out a problem that has not been solved by a formalist account of mental representation. It is the case independently of Searle's own account of intentionality (Searle 1983) that is based on his intuition about the role of consciousness.

2.4.2. The systems reply

The systems reply to Searle's experiment comes down to a claim that although Searle, as a person confined in the Chinese room, does not understand Chinese, the whole system does (Searle 1980). Supporters of the systems reply point out that the fact that a person in a room does not understand Chinese does not imply that the system does not. And Searle has never shown anywhere that a whole system does not understand it. According to his opponents, Searle makes a mistake of identifying the part of a system with the system as a whole. Searle would be right only if understanding could be divided like mass. We may cut an apple into pieces: the mass applies both to an apple and to its component parts. Searle must in fact assume that every part of a systems thinks. So, if a person himself understands, a stomach or a liver, for example, understands too; if a stomach would not understand, then a person would not understand too (Copeland 1993).

In response to this objection, Searle argues that, based on his opponents' arguments, a system that has memorized incomprehensible rules constitutes, together with a sheet of paper, a thinking system, which would be absurd. Searle states that there must be a difference between "genuinely mental" systems and those that are not genuine, and that the system itself must be able to detect the difference. Such a system displays - according to him - biologically "hardwired" intentionality.

The problem posed by Searle is deep but his solution unsatisfactory. This is because Searle *a priori* settles the intriguing problem of demarcation between thinking systems and other systems, without showing what the differences actually are. The claim that systems are different definitely does not suffice as a solution. And why exactly are consciousness and biological brains important according to him remains unexplained.

2.4.3. Chinese Room thought experiment reloaded

In this section, I will briefly present Searle's experiment in a version slightly modified by S. Harnad; here I also separate the problem of intentionality from the problem of consciousness. Harnad helps us better describe the problem of the relation between a formal symbol and reference of mental representations, which, thanks to his paper, in

artificial intelligence has been called "the symbol grounding problem" (Harnad 1990).

Harnad's experiment has two versions. In the first version, he describes learning Chinese as the first language using only a Chinese-Chinese dictionary, while in the second one, he describes learning Chinese as a second language. The latter one he considers to be feasible, though difficult. However, theoretically crucial is the first one.

In Harnad's variant, the Chinese Room experiment is about grounding symbols in something other than other meaningless symbols. It is not other symbols, according to Harnad, that constitute meaning, but relations of the cognitive system with the world. The problem is not whether it is possible to translate some specific language of a given linguistic form into another language of another linguistic form, but rather how to relate a linguistic form to the world.

Searle's argument, which is more evident in Harnad's version of it, is that syntactic properties are not sufficient to capture semantics, because for a proper description of semantics one also needs reference and logical value. The argument is thus directed against the internalism of a formalist vein. The formalist account does not allow us to account for the reference. As Fodor shows in the case of sentences with intensional contexts, formal properties of a vehicle make it possible to infer much about truth or falsity of representation. A formalist account does not, however, provide any explanation as to why a representation can be a vehicle of truth. Even if you accept a formalist assumption that mental representations have syntax, which is causally efficacious, it is not clear at all that representations have a property of being true of false. Moreover, under formalistic assumptions, it is by no means clear what physical structure could be considered syntactic and why certain syntactic constructs would correspond to falsehoods and others to truths.

2.5. Internalism and intension determination via learning

Internalists have difficulties in explaining action, which is related to their psychologically implausible approach to learning concepts. Fodor assumed that there were necessary and sufficient conditions for having concepts and that an account of narrow content should serve as a satisfactory psychological theory of learning (Fodor 1980). Such an

approach to learning does not allow, however, to take into account determination of intension via learning.

Burge's argument (Burge 1979) is aimed against the account of narrow content and is based on the human capacity to learn concepts. In his example, we are dealing with a thought experiment built analogously to Putnam's experiment, but Burge argues for a role of social context in content determination rather than for the determination by some physical facts. In the experiment, we compare a person whose physical states from birth until now are the same, but which occur in two situations that differ only in the linguistic community or in the social environment. As a consequence, such a person in these two situations uses a term *arthritis* differently: in the first situation the person knows well the extension of the term, and in the other, he or she uses the term *arthritis* to designate a disease that can occur in both muscles and joints. According to Burge, the extension of the term depends on the social context in which the person is raised and in which such a term is used. A defender of narrow content could answer that in the second situation the person has only an inkling about arthritis. He or she knows only that it is a disease but he or she is mistaken about what kind of disease it is. However, the defenders of narrow content go too far. They claim that assigning to a person any knowledge of the term's extension is unjustified in the second case, since it is not known what it describes.

Usually, learning concepts is time-consuming and gradual. However, according to the classical account, you have the same concept only if you mastered it completely; so the concept of arthritis that is not fully mastered is not yet a concept of arthritis. If that is the case, then we would have to assume that when we do not know necessary and sufficient conditions of concept application (and Fodor himself argues for the claim that, in general, we do *not* know them; cf (Fodor et al. 1980)), we do not know the same concepts. Without the assumption of innateness, this leads to a very peculiar consequence. It is not easy to indicate necessary and sufficient conditions for the use of terms such as "game", "chair", or "animal", and if they were not innate, then according to the classical theory of concepts, we should say that we do not know them at all. However, if they are innate, then their content is not determined individually. Here again we come across the fundamental aporia of the Fodorian account: his nativism excludes methodological

solipsism, since innate concepts must have content determined by factors that do not supervene locally, i.e., have content that goes beyond narrow content.

What an internalist, such as Fodor, argues against Burge, exposes the weaknesses of Fodor's internalism. Of course, an internalist could give up also the classic theory of concepts and nativism, but then Burge's argument would strike him. So he would have to agree that some determinants of content are social.

2.6. Summary

In this paper, I demonstrated that the argument of one-factor internalist account for the sufficiency of narrow content in the theory of representation is inadequate. To summarize, the characterization of narrow content leads either to ambiguity or to depriving the resulting concept of content of semantic properties. If by "narrow content" we mean – like Segal – the property of representational content that is completely independent of external factors to a cognitive system, the concept of content remains elusive and nobody knows what it could be. Although understanding content as partially independent from contextual factors allows us to hold content properties invariant in various situations, it seems that understanding content in total abstraction from the external factors of such properties does fit the bill. On the other hand, Fodor's formalistic account, in particular in Stich's radical interpretation, eliminates the properties of content to replace them with syntactic ones.

Reasons quoted by defenders of narrow content, such as the ability to articulate thoughts independently of the context or thinking about non-existent objects, speak in favor of the concept of narrow content. Nevertheless, the concept of narrow content abstracts away from both reference and satisfaction conditions, without which it is impossible to understand how mental representations can be vehicles of content. The lack of connection to reference and satisfaction conditions makes it for the account of narrow content impossible to state anything about the adequacy of representation with regard to their targets or referents. Some of these representations apply to an environment, which, in the correct account of content, would explain adaptive behaviors of animals as based on adequate representations of an environment, such as orientation in an environment (e.g., through

cognitive mapping). Narrow accounts of content do not allow us to state that, for example, a predator made a mistake in hunting while looking for a victim. For this reason, semantic internalism is a mistake as a solution to the problem of intentionality.

REFERENCES

Block, Ned. 1987. "Advertisement for a Semantics for Psychology." *Midwest Studies in Philosophy* 10 (1): 615–78. doi:10.1111/j.1475-4975.1987.tb00558.x.

Burge, Tyler. 1979. "Individualism and the Mental." *Midwest Studies in Philosophy* 4 (1): 73–121. doi:10.1111/j.1475-4975.1979.tb00374.x.

Carnap, R. 1947. *Meaning and Necessity*. Chicago: University of Chicago Press.

Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

———. 2004. "Epistemic Two-Dimensional Semantics." *Philosophical Studies* 118 (1/2): 153–226. doi:10.1023/B:PHIL.0000019546.17135.e0.

Copeland, B. Jack. 1993. *Artificial Intelligence: A Philosophical Introduction*. Oxford, UK; Cambridge, Mass.: Blackwell.

Copi, Irving, and Carl Cohen. 2002. *Introduction to Logic*. 11thed. Upper Saddle River N.J.: Prentice Hall.

Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge, Mass.: MIT Press.

———. 1991. *Consciousness Explained*. New York: Back Bay Books / Little Brown and Company.

Field, Hartry. 1978. "Mental Representation." *Erkenntnis* 13: 9–61.

Fodor, Jerry A. 1975. *The Language of Thought*. 1sted. New York: Thomas Y. Crowell Company.

———. 1980. "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology." *Behavioral and Brain Sciences* 3 (1): 63–63. doi:10.1017/S0140525X00001771.

———. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, Mass.: MIT Press.

———. 2008. *LOT 2: The Language of Thought Revisited*. Oxford University Press.

Fodor, Jerry A., M F Garrett, E C Walker, and C H Parkes. 1980. "Against Definitions." *Cognition* 8 (3): 263–367.

Griffiths, Paul Edmund. 2002. "What Is Innateness?" *The Monist* 85 (1).

Gulick, R. van. 2001. "Reduction, Emergence and Other Recent Options on the Mind/body Problem. A Philosophic Overview." *Journal of Consciousness Studies* 8 (9–10): 1–34.

Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D* 42: 335–46.

Hauser, Larry. 1997. "Searle's Chinese Box: Debunking the Chinese Room Argument." *Minds and Machines* 7 (2): 199–226.

Heath, J. 2015. "Methodological Individualism." In *The Stanford Encyclopedia of Philosophy*.

Jadacki, Jacek. 2001. *Spór O Granice Języka : Elementy Semiotyki Logicznej I Metodologii*. Warszawa: Wydawn. Naukowe Semper.

Maruszewski, Tomasz. 1983. *Analiza Procesów Poznawczych Jednostki W Świetle Idealizacyjnej Teorii Nauki*. Poznań: Wydawn. Nauk. Uniwersytetu im. A. Mickiewicza.

Pitt, David. 2013. "Mental Representation." Edited by Edward N Zalta. *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/fall2013/entries/mental-representation/.

Poczobut, Robert. 2007. "Eksternalizm Treści Umysłowej a Superweniencja." *Kognitywistyka I Media W Edukacji2* 9 (1): 82–106.

Preston, John, and Mark Bishop. 2002. *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*. Oxford; New York: Clarendon Press.

Putnam, Hilary. 1975. "The Meaning of Meaning." In *Philosophical Papers, Vol. II : Mind, Language, and Reality*. Cambridge: Cambridge University Press.

Samuels, Richard. 2002. "Nativism in Cognitive Science." *Mind and Language* 17 (3): 233–65. doi:10.1111/1468-0017.00197.

———. 2004. "Innateness in Cognitive Science." *Trends in Cognitive Sciences* 8 (3): 136–41. doi:10.1016/j.tics.2004.01.010.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 1–19. doi:10.1017/S0140525X00005756.

———. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge / New York: Cambridge University Press.

Segal, Gabriel. 2000. *A Slim Book about Narrow Content*. Cambridge Mass.: MIT Press. http://www.worldcat.org/title/slim-book-about-narrow-content/oclc/42771559&referer=brief_results.

Stich, Stephen P. 1980. "Paying the Price for Methodological Solipsism." *Behavioral and Brain Sciences* 3 (1): 97. doi:10.1017/S0140525X00002016.
Yablo, S. 1999. "Intrinsicness." *Philosophical Topics* 26 (1/2): 479–505.

ABSTRACT

SEMANTIC INTERNALISM IS A MISTAKE

The concept of narrow content is still under discussion in the debate over mental representation. In the paper, one-factor dimensional accounts of representation are analyzed, particularly the case of Fodor's methodological solipsism. In methodological solipsism, semantic properties of content are arguably eliminated in favor of syntactic ones. If "narrow content" means content properties independent of external factors to a system (as in Segal's view), the concept of content becomes elusive. Moreover, important conceptual problems with one-factor dimensional account are pointed out as a result of analysis arguments presented by J. Searle, S. Harnad and T. Burge. Furthermore, these problems are illustrated with psychological and ethological examples. Although understanding content as partially independent from contextual factors allows theorists to preserve content properties, it seems that understanding content in total abstraction from external factors of these properties is implausible. As a result, internalism is rejected in favor of externalism.

KEYWORDS: internalism; externalism; one-factor dimensional account of representation; mental representation; Fodor; methodological solipsism

HYBRIS nr 38 (2017)
ISSN: 1689-4286

MARIA MATUSZKIEWICZ
UNIVERSITY OF WARSAW

KNOWLEDGE ABOUT OUR EXPERIENCE AND DISTINGUISHING BETWEEN POSSIBILITIES

In his John Locke lectures delivered at the University of Oxford and published as a book *Our Knowledge of the Internal World* (2008), Robert Stalnaker characterizes the difference between two opposing philosophical perspectives: the externalist and the internalist one in terms of their starting points. For an internalist, it is mental contents accessible through introspection that are the foundation of knowledge. The philosophical question an internalist asks is *how can our knowledge reach beyond the contents of our mental states*? For an externalist, it is the external world with the objects, properties, and relations within it that are the starting point. Among these objects there are creatures who have thoughts and experiences. The question is: *how can these objects – human beings – have thoughts, which are about the world and about themselves?*

Stalnaker argues that puzzles concerning knowledge about our experience originate in the conflation of the two perspectives. As a remedy, he proposes a more thorough form of externalism. Externalism conceived in such a way consists in not only the claim that the contents of our mental states are determined by external facts (including both natural facts and the social environment), but also in the claim that contents are essentially ascribed. This type of externalism is linked to contextualism: contents are always ascribed in a particular context and there is no single correct characterization of our mental state, independent of the context of ascription. Stalnaker argues that providing a solution to some philosophical problems, problems of intentionality and of knowledge included, requires a shift from the perspective of a subject to the perspective of a theorist.

The main problem that the book addresses concerns our epistemic relation to our experience and the relation between experience and knowledge. Discussing Frank Jackson's knowledge argument and a solution to the puzzle it describes, Stalnaker argues that, contrary to what the empiricist tradition takes for granted, our knowledge about our experience is no more direct than our knowledge about external objects. Stalnaker's solution to the knowledge argument is based on the analogy between phenomenal and self-locating knowledge. Both are accessible only from a particular perspective. The most complex parts of the argument presented in the book concern the relation between the knowledge we can have only from a certain perspective and the objective knowledge.

The knowledge argument and the three strategies

Stalnaker starts with a discussion of the knowledge argument and three different strategies to avoid its conclusion.

The knowledge argument is based on two premises: (1) Mary knows all the facts of the type F. (2) Mary doesn't know the fact that *p*. From the two premises it follows that (3) The fact that *p* is not of the type F. The argument leads to the conclusion that beyond physical facts (or communicable facts) there exist facts of a different kind, phenomenal facts. Stalnaker discusses three strategies to resist the argument. The Fregean strategy adopts a more fine-grained notion of information than that of discriminating possibilities. David Lewis claims that what Mary acquires is not new information, but a new ability. John Perry's solution draws on an analogy between phenomenal and self-locating knowledge. All three strategies attempt to rebut the argument by ruling out the possibility that what Mary lacks is certain information, understood in terms of distinguishing between possibilities. It is this possibility that Stalnaker urges us to recognize.

i. The Fregean strategy

A proponent of the Fregean solution claims that Mary does not learn a new fact, but she learns the same fact in a new way. The solution requires modes of presentation or senses which individuate thoughts in a more fine-grained way than in terms of their truth conditions. Before leaving the black and white room, Mary has knowledge about certain mental state under a functional or neurophysiological mode of presentation. As a result, she knows the same fact under a visual mode of presentation.

Stalnaker rejects this solution, because, all in all, it does not avoid the conclusion that Mary's lack of knowledge stems from her inability to eliminate possibilities, and at the same time it does not provide a good account of these possibilities. The argument takes the following form: if materialism is true, both of these concepts (modes of presentations) necessarily pick out the same object. We might have both concepts and yet not know that they apply to the same object: no *a priori* reasoning leads from the one to the other. Hence, we might conceive of a situation in which a certain object is ϕ, without being ψ. We are forced to accept that the situation is epistemically possible, although it is not metaphysically possible. Stalnaker, however, rejects the idea that metaphysical possibilities are a proper subset of a broader class of possibilities, which include conceptual possibilities. His objection is that we have no conception of a merely conceptual possibility. Usually what is merely conceivable is defined in terms of what one may have a clear conception of. Stalnaker argues that we cannot have a clear conception of an impossibility. In cases of necessary *a posteriori* truths, one way of explaining modal illusions is to redescribe them: for example a situation where one thinks that water is not H_2O can be redescribed as a situation where one thinks of a substance phenomenally alike water that it is not H_2O. Phenomenal experiences do not – one might argue in Kripke's line of thought – allow for such a redescription. Let's suppose that having a red sensation is identical to a functional state F. Yet, I can think that it is not. The possibility that I conceive of – one would argue – cannot be adequately redescribed as a possibility that some experience, other than the sensation of red, is not a functional state F. It is because the phenomenal property (the property of being experienced as seeing red) is essential to seeing red. Stalnaker presents this view just to reject it. As the objection is a step in his argument against the Fregean solution, one might worry whether – since Stalnaker himself rejects this objection – he does not dismiss this strategy too fast.

ii. The ability hipothesis

Lewis (1988) rejected the assumption (2) that what Mary lacks is knowledge of a certain fact. What she acquires after leaving her room is a new ability, not knowledge. This ability is not of a cognitive kind. Cognitive abilities enable us to distinguish between possibilities. Lewis argues that neither Mary's situation before her release nor afterwards

can be described in terms of an ability to distinguish between possibilities. Lewis argues that before having an experience one cannot represent different possibilities. Before her release, Mary could not represent what it is to see red or what it is to see green and hence – according to Lewis – these possibilities cannot be used to characterize her mental state. After her release she can only think that *it is like that to see red* and again she cannot distinguish between different possibilities. Stalnaker points out that *post factum* Mary can represent different possibilities that she could not represent before and that these possibilities can be used to characterize her past cognitive limitations. What is important is that possibilities play an external role in characterizing the thinker's mental states: it is the theorist who uses them in order to ascribe mental contents. Stalnaker argues that the ability that Mary acquires is a cognitive ability. "While it may be right, as the ability hypothesis claims, that Mary does not necessarily acquire information merely by having color experience, it seems that she does acquire an ability to make distinctions between possibilities that she could not distinguish before, and a proper account of these abilities requires an account of the distinctions between the possibilities" (Stalnaker 2008:37). The quote is key to understanding Stalnaker's view on the relation between the knowledge about our experience and the knowledge about the world.

iii. The self-locating analogy

John Perry's strategy is, according to Stalnaker, the most promising one. It draws on the analogy between self-locating and phenomenal knowledge. The analogy suggests how to avoid the conclusion of the argument. While we might know all facts of type F and yet still lack some self-locating knowledge, very few philosophers would conclude that there are self-locating facts 'over and above' physical facts.

Perry's solution proposes that beliefs and utterances have more than one type of content: aside from subject-matter content, they also include reflexive contents (Perry, 2001). Subject-matter contents are conditions that the world has to satisfy for the belief or the utterance to be true. Reflexive contents are conditions not only on the world but also on the belief or the utterance itself. When John says "I am happy", the subject matter content of his utterance is *that John is happy*. Its reflexive content is *that the person having this thought token is happy*. In some

situations, all we learn when acquiring a self-locating belief is a reflexive type of content. For example, I know that the meeting starts at noon. When suddenly I realize that the meeting starts now, all I learn is a reflexive content: that the time of this thought token is noon. What I learn does not commit me to the idea that there are some self-locating facts over and above physical facts. Perry adopts this strategy to solve the puzzle about Mary: what Mary learns upon her release is a belief with a different reflexive content. Before her release Mary knew (1) *QR is what it is like to see red* (*QR* being a functional concept of seeing red). What she learns upon her release is (2) *this*$_{RED}$ *is what it is like to see red* (this$_{\text{RED}}$ being a visual concept of seeing red). The two thoughts have the same subject matter content, but they differ in their reflexive content: "(1) is true iff the origin of Mary's QR concept, the concept involved in (1), is the subjective character of the experience of seeing red". (2), on the other hand, "is true iff the act of inner attention to which it is attached is of the subjective character of the experience of seeing red" (Perry 2001: 147-148).

While Stalnaker accepts the general intuition behind the notion of reflexive content and agrees that an adequate theory of beliefs requires an account of how we can represent the perspective from which we perceive the world, he criticizes the way Perry has introduced the notion. First, he argues that Perry confuses the means of a representation with its contents. While the distinction between subject matter and reflexive content applies to utterances, its application to beliefs is problematic (Stalnaker 2008:39-40). Utterances are different than beliefs in that we might individuate them also in terms of occurrences of certain patterns of sounds, not only in terms of their content. With regard to beliefs, we cannot individuate them in any other way than in terms of their content[1], that is, in terms of the proposition they express. Thus, when we speak about different types of contents regarding beliefs, we must assume a vehicle of content: language of thoughts or inner symbols representing the content. Stalnaker insists that instead of distinguishing between different types of contents we should model all the relevant aspects of

[1] We might of course individuate a thought by referring to it as *Mary's favorite thought*. We may, however, still ask *what is her favorite thought?* In order to know whether Mary's favorite thought is the same as John's we need to know their contents.

content (including the thinker's perspective) in terms of distinguishing between possibilities.

It is worth noting that Stalnaker is generally critical with respect to theories which attempt to account for the intentional character of our thoughts by postulating concepts, mental files, and other inner vehicles of contents. His main charge is that these theories confuse intentional with non-intentional descriptions and they merely pretend to explain intentionality. For example, the mental file metaphor explains the difference between two beliefs with the same subject matter content, in terms of the difference in the mental files involved, which 'store' these contents. Stalnaker reminds us however that what these files supposedly consist of are not propositions, but certain physical objects, whose intentional properties still require explanation (Stalnaker 2008: 40).

Stalnaker also points out that reflexive contents as such do not explain Jackson's puzzle. He evokes Nida-Rümelin's thought experiment to demonstrate that the analogy between what we learn when we acquire a self-locating belief and what we learn when we have a phenomenal experience is flawed. In Nida-Rümelin's scenario, Mary's cognitive achievement is divided into two steps (Nida-Rümelin, 1995). Upon her release, Mary is first transported to a room covered with a multi-colored abstract wallpaper. Mary experiences colors for the first time, but she cannot connect her sensations to the concepts she has had before. It is only at the second step that Mary learns which color is which. It is then that she acquires information which is analogous to a self-locating belief. Thus, the analogy – Stalnaker concludes – cannot explain what she learns at the first step.

Stalnaker discusses yet another – although in his own view apparent – difficulty with the analogy. Many philosophers would say that there is an asymmetry between the two types of knowledge: while the information one learns when he acquires a self-locating belief is a contingent one (what I learn from you saying "I am Smith", can be redescribed as: "a person looking such-and-such is Smith"), the information one learns by having an experience is not contingent in that sense. In the first case, the argument goes, we are allowed to interpret the demonstrative reference descriptively, as a non-rigid designator. In the second case, "*this* is what it is like to see red", doing so would amount to saying that in a different possible world some other experience would play the role of seeing red. Many philosophers reject such a possibility.

Stalnaker in his critique of direct knowledge aims to show that they are wrong.

Stalnaker's account of self-locating beliefs

The analogy is helpful, according to Stalnaker, on the condition that we have an adequate theory of self-locating beliefs. The account Stalnaker proposes differs from Perry's and Lewis' (Lewis, 1979) in that he claims that we can explain the special nature of *de se* beliefs in terms of their content without introducing a special type of contents (reflexive contents) or without altogether modifying the notion of content (Lewis' centered worlds). The solution rests on the assumption that a lack of a self-locating knowledge always amounts to ignorance with respect to which of the worlds is the actual one. For the solution to work, it must provide such a mode of a transworld identification of the thinker, which excludes the possibility of the thinker's not knowing that the one who is thus identified is he himself. The mode of identification which satisfies this condition is by reference to one's occurrent thought token (Stalnaker 2008: 61).

Let's recall Lewis' example with two gods (Lewis, 1979): one god lives on the highest mountain and throws manna, the other god lives on the coldest mountain and throws thunderbolts. The gods are omniscient in the sense that they have all propositional knowledge. What they don't know is which one is which. According to the account proposed by Stalnaker, what each of the two gods doesn't know is which of the two worlds is the actual one: the world in which a person having *this* thought is a god living on the highest mountain or the world, in which a person having *this* thought is a god living on the coldest mountain. Stalnaker calls his solution a *haecceitistic* one, claiming that the worlds thus distinguished are qualitatively indiscernible (Stalnaker 2008: 58-59). This may give rise to three concerns: (1) whether the identification by reference to an occurrent thought token is really immune to the error through misidentification (that is, whether it is such that the thinker cannot be unaware that it is him who is thus identified) (2) that the solution commits us to existence of possibilities accessible only from a first-person perspective; what is the relation between the self-locating knowledge and the objective knowledge? (3) are the differences between the two worlds really merely *haecceitistic*, and do we have good reasons to think that some possibilities do not differ qualitatively?

Essentially indexical information and the relation between it and the objective information

Stalnaker emphasizes that our theory of beliefs requires a notion of informational content which can be separated from its relation with the thinker whose knowledge it represents. We may have complete objective knowledge about the world, yet lack some self-locating information. Is then the self-locating information something "over and above" the objective information?

Not many philosophers are likely to draw this conclusion, unlike in the case of phenomenal knowledge. The analogy between the two kinds of knowledge is that in both cases the epistemic situation of the subject is represented by possibilities which can be distinguished only from a particular perspective. We need to explain the relation between the self-locating knowledge and the objective knowledge.

A notion important to Stalnaker's view is the notion of *essentially indexical information*. Essentially indexical information consists of "distinctions between the possibilities (the ways the world might be) that can be represented only from a certain perspective, but that once represented, can be abstracted from the perspective" (Stalnaker 2008:78). We might explain the notion using one of the examples discussed by Stalnaker. Sleeping Beauty (the heroine of Adam Elga's puzzle), before being put to sleep learns that she will be woken up once (on Monday) or twice (both on Monday and Tuesday) depending on the result of a coin toss (Elga, 2000).

> On Monday (and Tuesday, should she be awakened then), Sleeping Beauty was able to distinguish between a world in which, as she would put it then, *today* is Monday, and a different world in which *today* is Tuesday. On Sunday she was unable to distinguish between these two possible worlds, since in both of them an event of the same kind occurred on both Monday and Tuesday. To distinguish one from the other, one had to be there, or alternatively, to remember later having been there: one had to be in a position to refer uniquely to *that* particular time that Sleeping Beauty was awakened. But even on Sunday, Beauty was able to *describe* the distinction she was unable to make (Stalnaker 2008:78).

The notion of essentially indexical information is best understood in the light of Stalnaker's theory of communication. Communication always takes place in a context (which includes beliefs held by the participants in the conversation) and results in a change of context. The context is

best represented as a set of possibilities (possible worlds). While communicating, the participants change the context by adding new information i.e. by excluding some possibilities. Sometimes knowledge of the relation between the utterance and the context is important to determine what information is being communicated. It is, however, not itself part of that information. For example, when I tell my friends "I live in Warsaw", they can extract the information from the context in which I communicated it. On the other hand, when I introduce myself by telling you my name, the information I thus convey cannot be extracted from the context of utterance, because "there is not a piece of information that is the content of what I told you that you can simply add to your stock of beliefs about the objective world". As Stalnaker explains, "the point about essentially contextual information is that sometimes the content of what is expressed or believed in is not detachable from the context in which it is expressed or believed" (Stalnaker, 2008: 81).

Whether one represents the thinker's beliefs by means of locally-distinguishable possibilities or by means of non-local possibilities depends on his (the theorist's) goal. Stalnaker illustrates this kind of context-dependence of belief ascriptions with the following example (Stalnaker 2008: 83-84): imagine Rudolf Lingens, the famous amnesiac from the Stanford Library. His two colleagues Daniels and O'Leary, who don't know his true identity, call him "Nathan". One day they see a crowd of journalists gathering in front of the library. O'Leary asks his friend: "Do you know who it is?" pointing at a man surrounded by journalists. "Yes" – Daniels replies – "it is our famous amnesiac friend, Nathan". In this context – as Stalnaker argues – we can represent Daniels' beliefs straightforwardly as beliefs about Lingens: the possibilities he eliminates are those in which it is someone other than Lingens whom they see. On the other hand, if one of the journalists approaches Daniels and asks him "Do you know who it is?" pointing at Lingens, the theorist would like to emphasize Daniels' ignorance with respect to his friend's true identity and he would represent Daniels' epistemic situation using locally-distinguishable possibilities: *the man who is there is X, Y, Z etc.* What is local about this characterization of Daniels' beliefs is that it is only relative to this context that we cannot describe Daniels' beliefs in terms of Lingens himself and that the possibilities that we use to characterize his beliefs are distinguishable only within that particular context.

Distinguishing vs. eliminating possibilities, and possibilities accessible only from a first-person perspective

Before I move on, I want to raise some concerns regarding the proposed account of self-locating knowledge. Key to Stalnaker's theory of beliefs (including self-locating beliefs) are notions of distinguishing and eliminating possibilities. It is important not to confuse these two notions.

To distinguish between possibilities is to be able to represent them (descriptively or by means of individuals), and it does not require one to know which possibility is the actual one. When I tossed a coin, before I check the result, I distinguish between two possibilities, although I cannot tell which one corresponds to the actual world. To eliminate possibilities, on the other hand, is to know which one of them is actual. The distinction between distinguishing and eliminating possibilities carries on to possibilities accessible only from the first-person perspective.

One problem related to Stalnaker's theory of *de se* beliefs is that it commits us to the view that some possibilities are accessible only from the first-person perspective. Stalnaker's solution consists in proposing a mode of identification of the thinker in terms of his occurrent thought token. This requires that we cannot have knowledge about particular thought-token other than from a first-person perspective. The claim doesn't seem controversial, but if we assume the possibility that thoughts are token-identical with physical events, it is less obvious why we cannot in principle have singular thoughts about someone else's thought tokens. If we could, we face again the possibility of error through misidentification: My belief "the person who is having *this* thought is X" does not imply a belief „I am X". We either have to rule out that thoughts are token-identical with physical events or we need to claim that there is a class of physical events which are accessible only from the first person perspective.

Second problem is how to reconcile the two claims that Stalnaker accepts: (1) a complete objective knowledge about the world requires the capacity to eliminate all possibilities which are inconsistent with the way the world actually is (i.e. one has to know the truth value of every proposition); (2) having a complete objective knowledge doesn't require one to have the capacity to eliminate these subjective possibilities (i.e. there are some propositions whose truth value one doesn't know).

One might answer that possibilities which we can represent only from the first-person perspective differ ontologically from other possibilities in that they are not real possibilities but mere representations of real possibilities. This, however, is at odds with Stalnaker's conception of possible worlds (which are not representations but real possibilities for the world, i.e. ways the world might be). Possibilities that the thinker can distinguish only from the first-person perspective should also be understood as real possibilities.

Second response one might offer is that possibilities which one cannot represent are irrelevant to one's knowledge and hence cannot be used to characterize one's epistemic situation. This is simply not true. The thinker need not have the ability to represent nor eliminate possibilities for these possibilities to be used in characterizing his beliefs. Suppose I am not aware of the existence of Plato. I cannot distinguish between nor eliminate the possibility in which Plato wrote *The Republic* and the possibility in which he didn't. It doesn't follow that this is irrelevant to my knowledge. To the contrary, we will characterize my ignorance in terms of these possibilities.

Finally, one might point out that knowledge ascriptions are context-dependent. Not being able to eliminate some possibilities in most contexts does not preclude knowledge ascription. Thus, when we say "X has all the propositional knowledge, but doesn't know whom he is" (as is the case of Lewis' gods), we are restricting the quantifier. We are leaving aside these propositions which are accessible only from the first-person perspective. This answer allows us to reconcile the two claims at the cost of a commitment to possibilities that are accessible only from the first-person perspective. Moreover, we have to accept their existence as a primitive fact.

The puzzle about Mary and the self-locating analogy

How does the self-locating analogy help to solve the puzzle about Mary? In both cases – of not knowing who we are and of not knowing what it is like to experience something – the information one lacks is essentially indexical in the sense that it distinguishes between possibilities representable only from a local perspective. When we acquire a self-locating belief, we eliminate possibilities that we could represent only from a local point of view. When we learn *this is what it is like to see red*, do we **likewise** come to **eliminate** possibilities? Stalnaker answers in

the negative: he claims that just by having a phenomenal experience one acquires a cognitive ability which enables one to represent possibilities but not to eliminate them. Stalnaker's argument supporting this claim rests on the assumption that we don't gain knowledge about its essential properties just by having a phenomenal experience. It is, however, not clear how that establishes that mere knowledge about our experience does not enable us to eliminate possibilities, since: (i) In some contexts it seems natural to say that just having the experience enables one to eliminate possibilities; (ii) Stalnaker rejects the view that the thinker must know the essential properties of an object to have singular beliefs about it. Thus, in both cases (phenomenal knowledge and knowledge about external objects) whether we ascribe knowledge i.e. the ability to eliminate possibilities depends on a context. Stalnaker's claim must be weaker: just having the experience does not automatically, and in every context, amount to having the ability to eliminate possibilities.

Stalnaker presents the following thought-experiment (Stalnaker 2008:86) to support his claim that the experience itself does not enable us to eliminate possibilities: Mary is told before her release that she will be subjected to an experiment. Depending on a result of a coin toss she will be shown a red or green star. Before the experiment takes place Mary can represent two possibilities, none of which she can eliminate. After the experiment, in which she was in fact shown a red star, she still distinguishes between two possibilities, none of which she can eliminate. Mary thinks: *I know how it is to see red or how it is to see green. I don't know, which of the two colors I saw*. What has changed about Mary's epistemic situation, according to Stalnaker, is that after the experiment she is able to represent knowledge about her own experience that she couldn't represent before. She is not, however, able to eliminate possibilities. The argument doesn't seem to be conclusive: one might argue that in some contexts it is intuitive to say that the experience does enable her to eliminate possibilities (I discuss such an example below). The conception of knowledge about our experience should make sense of such cases.

Lewis' theory of knowledge and the principle of phenomenal indistinguishability

Stalnaker provides an insightful critique of Lewis' theory of knowledge. He blames the inconsistencies of Lewis' view on him conflating the

externalist and internalist perspectives. On one hand, Lewis imposes very strong epistemic constraints on knowledge of objects, which makes him deny that we can have singular beliefs about them. On the other hand, he grants experience a role that it cannot play. That is because unless we accept the controversial claim that by merely having an experience we know its essential properties, our knowledge about our experience does not satisfy Lewis' restrictive criteria.

The principle of epistemic indistinguishability is the claim that worlds which are epistemically accessible to a thinker are phenomenally indistinguishable. Stalnaker defines the notion of phenomenal indistinguishability in terms of a cognitive capacity: two mental states are phenomenally indistinguishable iff the subject can switch from one to the other without noticing any difference. Stalnaker claims that the thought experiment with Mary and the two stars shows that the principle is false. After being presented with the red star, Mary still doesn't know which one is the actual world: the one in which she was shown a red star, or the one in which she was shown a green star. The two worlds should be phenomenally indistinguishable for her, while in fact they are not. If they were, we would have to accept that there is a counterfactual world in which she saw a green star, which is phenomenally indistinguishable from the actual one in which she saw a red star. The principle of epistemic indistinguishability commits us – Stalnaker concludes – to the existence of phenomenal information (Stalnaker 2008: 90-91).

We may, however, disagree with Stalnaker in that Mary doesn't know whether she was shown a red or green star. She knows which star she was shown, she only doesn't know the name of its color. If she was first shown a red star, and a moment later a green, yellow, and a blue one, and if she was asked which of the stars she saw first, she would be able to eliminate the possibilities. Stalnaker does indeed discuss a similar case: he claims that even when Mary names all the colors, but she cannot relate these names with the names she was using while locked in the black and white room, it is a matter of context whether we would ascribe to her knowledge that the object is red. In some situations of this type we tend to say that Mary knows that this tomato is red, and in another we don't (when we want to emphasize the fact that she is unable to connect her old concepts to her new experiences). Stalnaker, however, thinks that the fact that in some contexts we would be reluctant to ascribe

knowledge is enough to undermine the principle of phenomenal indistinguishability.

Stalnaker presents us with the following choice: either we stick to the principle of phenomenal indistinguishability at the cost of accepting that phenomenal information exists or we reject the principle and accept that knowledge about our experience is not epistemically privileged and does not play the role it was granted by empiricist epistemology. Stalnaker argues for the latter option: the knowledge about our experience is as indirect as knowledge about external objects. What he attempts to do, however, is to elucidate the notion of direct knowledge by explaining the intuitions that motivate it.

Lewis imposes a very strong epistemic constraint on the knowledge of objects: singular thoughts about objects require knowledge of their essential properties. Since we don't know essential properties of objects, Lewis claims that we cannot have singular thoughts about them. He uses Saul Kripke's puzzle about Pierre to justify this claim (Lewis, 1981). Pierre, as we remember, thinks about London (when he is still in France and calls it "Londres") that it is pretty and (when moving to London and using its English name) that it is not pretty. If we accept the theory of direct reference and the disquotational principle we are forced to ascribe to him contradictory beliefs. That violates our intuition that Pierre is rational. Lewis argues that the puzzle lends support to the internalist theory of beliefs. He argues that we might conceive of a situation in which the French name "Londres" designates a different city than London, Bristol for instance. Such a world is for Pierre indistinguishable from the actual one. Since Pierre doesn't know London's essential properties, there are such possible worlds epistemically accessible to him in which the name "Londres" refers to a different city than London. Whenever we have beliefs about objects whose essential properties we don't know, there are epistemically accessible worlds in which some other object plays the same role as the given object plays in the actual world.

An anti-individualist (an externalist) would object that the belief Pierre would have in that counterfactual world differs from the one he has in the actual world. Lewis rejects this counterargument claiming that we need a narrow notion of content in order to explain how we can have access to our own beliefs and to avoid the conclusion that Pierre is irrational.

Lewis claims that we cannot have a singular thought about an object without knowing its essential properties, and the same goes for our knowledge about phenomenal experience: we cannot have singular beliefs about our experience without knowing its essential nature. Lewis thus rejects the controversial claim that we know the essential properties of a phenomenal experience merely by having this experience. Accepting this controversial claim amounts to saying that by merely having an experience we would eliminate all possibilities in which my phenomenal experience has different physical nature. As Stalnaker points out, a materialist cannot accept this claim, as he holds that experiences are identical to physical states and that it is physical properties that are essential to them. We don't know the physical nature of our experiences merely by having them. Having a phenomenal experience of a given type, for instance a headache, is not sufficient. I do not know whether my experience is a complex physical state of type A or B.

Stalnaker points out that the privileged role that experience plays in Lewis' theory of knowledge is at odds with his rejection of the claim that just by having an experience one knows its essential properties (Stalnaker 2008: 99). According to Lewis' theory, knowledge is represented by possibilities which are not eliminated by experience (Lewis, 1996). A possible world *w* is not eliminated by experience iff the subject's perceptual experience and memory in *w* are the same as they are in the actual world. Lewis gives a contextualist response to a skeptic's concern "how then can we have knowledge which goes beyond our experience?": depending on the context, we are allowed to ignore some possibilities which our experience does not eliminate. Let's illustrate this idea with an example. John knows that it rains iff his experience eliminates every possibility in which it does not rain. We can, however, imagine that what John takes to be rain is an effect produced by a film crew. Although John's experience does not eliminate this possibility, in a normal context (when it actually rains) this does not preclude us from ascribing knowledge to him.

However, what is of key importance to Lewis' theory is the assumption that we will know that the possibilities eliminated by experience, in any context, are inconsistent with our knowledge. This – Stalnaker argues – implies that by having an experience we gain knowledge about its essential properties and hence implies the above

mentioned thesis (that having an experience is knowing its essential properties), which Lewis in fact rejects. Accepting this controversial thesis, as Stalnaker reminds us, implies the phenomenal indistinguishability principle together with its correlate, the existence of phenomenal information.

Stalnaker blames these inconsistencies in Lewis' theory on him conflating the externalist and the internalist perspectives. The theory is externalist in that it describes the experience from an external perspective: as a set of possibilities in which the subject's experiences are identical. It is, however, internalist in the privileged role of experience (Stalnaker 2008: 101). Lewis' theory identifies knowledge with possibilities not eliminated by experience. As Stalnaker points out, according to Lewis' theory, it is the mere occurrence of an experience that eliminates possibilities and not its propositional content. Worlds which are not eliminated by experience are worlds in which the subject has the same experiences as in the actual world. What the theory assumes is that two identical experiences have identical causes, hence there is a one-to-one relationship between phenomenal properties of our experience and the physical features of the world which cause this experience.

Stalnaker is critical of these ideas. He thinks that Lewis misconceives the role of experience in eliminating possibilities by not recognizing that whether we'll say that a certain experience eliminates possibilities or not depends on a context. Let's think once again about Stalnaker's thought experiment with Mary seeing a red or a green star. After being shown a red star, Mary is still not able to eliminate one of the two possibilities. But whether we'll judge her to know which star she has seen (and ascribe her the ability to eliminate possibilities) depends on our, the theorist's aims. Imagine two scenarios: in the first scenario we tell Mary that first she will be shown a red or a green star and later she will be presented with two buttons, a red and green one. If she presses the red button, the world will be annihilated. If she presses the green one, nothing will happen. In the second scenario, we tell her that later she will be presented with two buttons, a red and green one. If she presses the button of the color that she was exposed to during the experiment, the world will be annihilated. In the first case, we may say that the mere experience did not enable Mary to eliminate possibilities. In the second case, we will say that having the experience she has eliminated a

possibility. Thus, knowledge we ascribe on the basis of the experience depends on the context of ascription.

Lewis' contextualism is – according to Stalnaker – not thorough enough. Although Lewis thinks that in our everyday practice of belief ascription we are entitled to eliminate certain possibilities which seem irrelevant, we might still speak about knowledge in the absolute sense. Stalnaker disagrees with Lewis on this point. He also argues that it is not the mere occurrence of an experience which eliminates possibilities but their propositional content. As he points out, there is no direct connection between having an experience and the propositional content of our beliefs.

The principle of epistemic transparency and anti-individualism

Many theorist think, like Bertrand Russell, that having a singular belief about an object requires that the thinker is in a special epistemic relation with the object. Without our being acquainted with the object, the object cannot be the content of our beliefs. There is no single interpretation of what the relation of acquaintance amounts to. One possible interpretation of this notion was proposed by Lewis: in order to be acquainted with an object one has to know its essence (Lewis, 1981). Since we don't know the essential properties of objects and persons, a proponent of this view has to deny that we have singular beliefs about them. The problem generalizes our knowledge of properties and relations. Stalnaker demonstrates that, contrary to the traditional view, we are not in such a privileged relation with our experience.

The key idea to Stalnaker's theory of belief ascriptions is his "deep contextualism". It is the view that there is no one correct context-independent characterization of the thinker's beliefs. It is not the case that all context-dependent characterizations should be regarded as a mere approximation which could after all be substituted with correct context-independent characterizations (which we don't do for practical reasons). Stalnaker claims that contents are essentially ascribed and not inherent. On this theory, having singular thoughts is not a matter of a thinker's acquaintance with an object or his having a particularly rich conception of the object. Whether we characterize one's beliefs in terms of the object singularly or descriptively depends on the context of the ascription. Stalnaker demonstrates that we might have a very detailed conception of an object and yet not be aware that this conception refers

to one and the same object. In order to characterize such a belief, we'll need to do it descriptively. On the other hand, in a different context, one might know very little about an object beyond some contingent fact, and still we might be able to ascribe to him a singular belief about this object. It is the context of ascription that is critical to what we'll judge as a correct characterization of someone's beliefs.

Deep contextualism also provides a way to accommodate a popular intuition that for our thoughts to explain the agents' actions they must be epistemically accessible to them. Many authors have claimed (for example Paul Boghossian) that the intuition cannot be reconciled with anti-individualism, which holds that it is facts about the environment which determine the contents of our thoughts. Stalnaker thinks that an anti-individualist can make sense of the former intuition if he accepts the view that contents are externally ascribed by the theorist.

The principle of transparency, which expresses the above mentioned intuition, says that for two thoughts of a subject to have the same content the thinker must know *a priori* that it is so (likewise if his two thoughts have different contents, the thinker must know that it is so) (Boghossian, 1994). The conflict between anti-individualism and the principle can be seen in the case of contradictory beliefs: situations in which we hold contradictory beliefs about an object, not realizing that our thoughts refer to one and the same object or situations in which we have beliefs about an object, not being aware that our beliefs refer to two different things. In these situations we cannot ascribe beliefs in a way prescribed by the anti-individualist theory or we would have to conclude that the subject is irrational.

One type of arguments against anti-individualism makes use of the so called 'slow-switching scenario' (Boghossian, 1994). It is a thought experiment which involves a thinker being transported from one context to another (from Earth to Twin Earth) in such a way that he is not aware of the change and the two contexts are indistinguishable for him. The thinker's beliefs on Earth are about water. When he thinks, shortly after being transported to Twin Earth, "There is water in this lake", according to an anti-individualist, he has a false belief, which concerns water and not a true belief concerning XYZ. At the same time, we have a strong intuition that after years spent on Twin Earth, when having this type of thought, he no longer thinks about water, but about XYZ. He can also – to the detriment of our theories – compare his earlier beliefs with his recent

ones: while living for many years on Twin Earth he might recall: *once* (thinking of a specific episode from his Earthly life) *water tasted much better than nowadays.*

Thought experiments of this kind demonstrate that our intuitions concerning the contents ascribed change with the situation. The challenge they are meant to pose for an anti-individualist is that he should account for the mysterious change that the subject undergoes which makes his thoughts change reference. Stalnaker rejects this way of putting the problem as deeply misleading (Stalnaker 2008: 121). Instead of situating the change in the subject (in his head), we ought to explain our intuition that the reference of his thoughts changes in terms of the context in which we make these belief ascriptions. Generally, we ascribe beliefs to explain and predict the agents' behavior. It follows that we must ascribe beliefs in a way which is consistent with the assumption that the agent is a rational being. The principle of transparency reflects a different requirement: not on the subject, but on the theorist: the ascriptions that the latter makes cannot violate the subject's rationality assumption.

Stalnaker convincingly shows that the principle of epistemic transparency can be reconciled with externalism, understood however, I would argue, slightly differently than the early formulations of the view (Burge, 1979) suggest. Burge claimed about two thinkers, who (i) live in two linguistic communities that differ only in their use of the term "arthritis" and (ii) nevertheless associate the same set of descriptions with the term arthritis, that they do not have any common beliefs about the disease. Externalism, as it is interpreted by Stalnaker, does not exclude the possibility that in some particular context it might be perfectly fine to characterize their beliefs with the same sets of possibilities. Stalnaker's externalism claims that (1) the content of our mental states depends on the facts concerning our environment in the sense that we should look for an explanation of why we have beliefs of that content in our causal relations with the environment. He also states that (2) contents are essentially ascribed and not inherent and that these ascriptions are made in a particular context which includes the cognitive aims of the theorist who does the ascription. The theory is also externalist in the sense that when making ascriptions we use the resources which are available to us: that is objects, properties, and relations which are there in the actual world.

Stalnaker's book makes a strong case for externalism understood as a methodology rather than a metaphysical view. At the same time, he acknowledges important internalist intuitions (e.g. that we perceive the world from a certain point of view,) and shows that, by reversing the order of explanation, we can do justice to them on externalist grounds. Stalnaker shows that the possible worlds representation of content enables us to represent the subjective point of view of the thinker as well as the relation between his perspective and the way the world is in itself. The book is both very rewarding and very demanding. For, although Stalnaker avoids technical details, he connects variety of philosophical issues, often shifting the grounds of the discussion.

Acknowledgments
I am grateful to Tadeusz Ciecierski for his comments on the earlier draft of this paper.

REFERENCES

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

Elga, A. (2000), "Self-Locating Belief and the Sleeping Beauty Problem", *Analysis* 60:143-147.

Lewis, D. (1979), "Attitudes *de dicto* and *de se*", *Philosophical Review* 88: 513-543.

Lewis, D. (1981), "What Puzzling Pierre Does not Believe", *Australasian Journal of Philosophy* 59: 283-289.

Lewis, D. (1988), "What Experience Teaches", in Lewis, D. (1999) *Papers in Metaphysics and Epistemology*, Cambridge: CUP.

Lewis, D. (1996), "Elusive Knowledge", *Australasian Journal of Philosophy* 74: 594-567.

Nida-Rümelin, M. (1995), "What Mary Couldn't Know: Belief and Phenomenal States", in Metzinger, T. (ed.), *Conscious Experience*, Exeter: Imprint Academic: 219-241.

Perry, J. (2001) *Knowledge, Possibility and Consciousness*, Cambridge, Mass.: MIT Press.

Stalnaker, R. (2008) *Our Knowledge of the External World*, Oxford: OUP.

ABSTRACT

KNOWLEDGE ABOUT OUR EXPERIENCE AND DISTINGUISHING BETWEEN POSSIBILITIES

In my article I reconstruct the main threads of Robert Stalnaker's book *Our Knowledge of the Internal World*, which focuses on the problem of our epistemic relation to our experience and the relation between experience and knowledge. First, the book proposes an interesting view of externalism, which combines classical externalist claims with a contextualist approach to content ascriptions. The approach accommodates some important internalist intuitions by showing how content ascriptions can be sensitive to the perspective from which a subject perceives the world. Second, Stalnaker proposes a theory of self-locating and phenomenal knowledge, which should be understood in terms of differentiating between real possibilities. The puzzling upshot of this elegant solution is that it commits one to the existence of possibilities accessible only from the first-person perspective. Finally, Stalnaker presents an argument which shows that our knowledge about our phenomenal experience is no more direct than the knowledge about external objects. Stalnaker's claim that by merely having an experience we don't learn any new information seems, however, too strict in light of his contextualist approach to content ascriptions.

KEYWORDS: Robert Stalnaker; externalism; contextualism; phenomenal experience; self-locating beliefs