Polskie Towarzystwo Semiotyczne

STUDIA SEMIOTYCZNE

Tom XXXV • nr 1

PÓŁROCZNIK

COGNITION. CONSCIOUSNESS. LANGUAGE

Warszawa • 2021

Założyciel „Studiów Semiotycznych” (*Founding Editor*):

Jerzy Pelc
(1924–2017)

Zespół redakcyjny (*Editorial Board*):
Andrzej Biłat (*Editor-in-Chief*)
Dominik Dziedzic (*Assistant Editor*)

Redaktor numeru (*Issue Editor*):
Zbysław Muszyński

Rada naukowa (*Advisory Board*):
Jerzy Bartmiński (Uniwersytet Marii Curie-Skłodowskiej), Paul Bouissac (University of Toronto), Andrzej Bronk (Katolicki Uniwersytet Lubelski Jana Pawła II), Roman Kossak (City University of New York), Stanisław Krajewski (Uniwersytet Warszawski), Idalia Kurcz (SWPS Uniwersytet Humanistycznospołeczny), Witold Marciszewski (Uniwersytet w Białymstoku, Fundacja na Rzecz Informatyki, Logiki i Matematyki), Genoveva Martí (ICREA & Universitat de Barcelona), Adam Nowaczyk (Uniwersytet Łódzki), Stefano Predelli (University of Nottingham), Mieczysław Omyła (Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie), Piotr Stalmaszczyk (Uniwersytet Łódzki), Anna Wierzbicka (Australian National University), André Wlodarczyk (Université Paris-Sorbonne), Jan Woleński (Uniwersytet Jagielloński, Wyższa Szkoła Informatyki i Zarządzania)

Redakcja językowa:
Martin Hinton

Skład elektroniczny:
Dominik Dziedzic

Adres redakcji: Krakowskie Przedmieście 3, 00-047 Warszawa
e-mail: studiasemiotyczne@pts.edu.pl
http://studiasemiotyczne.pts.edu.pl/

ISSN 0137-6608; e-ISSN 2544-073X

© Copyright by Polskie Towarzystwo Semiotyczne

CONTENT

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXV, nr 1 (2021), s. 5–8
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxv1.01

From the Editor

ZBYSŁAW MUSZYŃSKI [*]

INTRODUCTION: MANY FACES OF REPRESENTATIONALISM[1]

Representation and Cognitive Semiotics

The subject matter of all texts comprising this volume is a category of representation. Although it is not always explicit, the reference to the notion of representation enables to bring together shared characteristics of research into consciousness, enhancement of cognitive processes, metaphor and modes of coding of information in the mind.

The category of representation brings research closer to semiotics. Representation is a basic theoretical category in cognitive science and in semiotics since the approaches of both relate to: "something that stands in for something else under a determined aspect" which corresponds to the definition of sign by Charles Sanders Peirce. The convergence of both sciences is possible due to the formation of a new, common field of research called "cognitive semiotics" whose objective is to integrate perspectives, methods and insight from cognitive science into the broader context of cognitive and neurobiological processes.

Cognitive semiotics study mechanisms and processes of meaning-making in all domains: the natural, the social, the cultural, in language and other sign vehicles, especially in perception, and in action. In classical cognitive science the notions of sign, language and mind are linked with studies on representation that is why studies on cognitive semiotics aim to incorporate the results of other sciences, using methods ranging from conceptual and textual analysis as well as experimental and ethnographic investigations (e.g., Daddesio, 1994; Zlatev, 2012; Konderak, 2013).

[*] Maria Curie Skłodowska University, Department of Philosophy. E-mail: zmuszyn@gmail.com. ORCID: 0000–0003–2534–7660.

[1] Translated by Ewa Muszyńska-Faizi.

Cognitive semiotics is the trans-disciplinary study of language, communication, media and mind. In cognitive semiotics, both phenomenological analysis and empirical methods are used. The goal is to produce a new approach to interrelations between different codes of communication such as language, gestures and pictures.

All states of mind, if they have content or are of informative character are representational states (representations in short) and refer to something else other than themselves. Such representations include neural states of digital and analogue character, of linguistic and non-linguistic (perceptional) character, of index and metaphorical character.

Representation and Knowing/Knowledge

Knowledge is a type of representation and is a product of cognitive skills of a subject. Not always is an individual conscious of processes generating this particular type of representation.

There is a problem of control over arising representations or knowledge when the mind of an individual is enhanced by artifacts. Do the beliefs forming in such a way meet the criteria required for knowledge? Not every representation acquires the state of knowledge (it is the case with sensual representations). Representations always should be produced by cognitive states of an individual. In the case of brain-computer interface systems, representations can form artificially circumventing natural cognitive processes.

There are problems of a loss of identity ("former self") through the interface with a cognitive artifact, as well as a loss of control over decisions of the hybrid cognitive system and its results. It is an effect of interactions between brain structures and forming representations linked to the actions of an individual and having impact on their cognitive and emotional dispositions, mental abilities and personal inclinations. However, the epistemological status of hybrid cognitive systems may be assessed, it can be accepted that their cognitive results can be cognitively complete.

Representation and Consciousness

Consciousness can be concisely characterised as an individualised state of information which is a functional representation. The notion of information is superior to the notion of consciousness in the sense that all states of consciousness are states of information. States of consciousness can be treated as a subgroup of information states. Since it is difficult to discuss information without a generally understood reference, it can be assumed that conscious states of information are forms of representation and they represent functions and actions. Referentiality is understood broadly and—as opposed to intentionality—it can be linked to early stages of information processing in the brain. Such information-bearing-states of consciousness are based on the assumption that for a system-organism to be aware of something, it has to somehow identify this "something" and represent it as "this something". Information has to be understood naturalistically as a state of a specific system-organism that can be differentiated from other functional states of representation of this system. Information held by a given organism is unique, as its form, meaning and functions, which can be performed in actions, have been formed by many unique developmental factors (individual and species-related) as well as environmental factors. Information states of a particular organism undergo individuation which leads to the creation of first-person (subjective) perspective. Therefore, information as a form of functional representation becomes virtually unique at a phenotype level.

Representation and Indexical "I"

The use of the notions of representation or information in semantic analyses of indexical "I" allows for making clearer the distinction between the user and the producer of linguistic tokens of the so-called pure indexicals, especially the tokens of "I". Both concepts—the user and the producer—have an intentional character which links them to the contents of mental states understood as representations. Semantics of these expressions is connected with philosophy of mind, and also—in naturalised version—with cognitive science and through this with the theory of representation. This relationship is not always visible as A. J. Jacobson observes: "I claim that philosophy of mind has woefully neglected a sense of 'representation' that is present in neuroscience and that is important" (2003, p. 190). Therefore, in his opinion for research in cognitive science, and in particular in neuroscience, it would be more appropriate to use "representation" in the meaning of "token-realization". Token-representation is in contrast to intentional-representation. This differentiation is significant for the analysis of texts and implies the possibility to naturalise representation in such an interpretation. Token-realizations are actual states of reality on biological and physical levels (without the need to be reduced to them), because as Jacobson notes: "Token-realizations are not about [...]; they are of their types in the sense of being instances of their type [...]" (2003, p. 191).

Representation and the Organization of Information

The format of mental representation—the external or internal vehicle of representation—is the way information is organized in the mind. Differences between formats of representations are understood in terms of differences in information processing. The type of vehicle of representation is connected with the proper mechanism of how iconic or discursive information is processed in the mind. These mechanisms also depend on the modality of representation. Another difference between the formats of representations concerns their predictive functions. The format of representation includes problems of representational primitives and the rules of information processing.

Representation and Metaphor

Metaphorical expressions can be understood as linguistic representations based on analogies generating metonymic series which reflect associations founded on codes, cultural contexts and subjective experiences. These metaphorical representations are different for both the interpreter and the creator of the metaphorical text. To search for the foundations of a metaphor it is necessary to focus on such methods of analysis which are proper to use for those who embody the metaphor in a text and for the interpreter. Perceiving a metaphor as a type of representation allows up to explain how a metaphor is generated and define methods for its analysis. Such an approach to metaphor opens up new facets of understanding and studying the phenomenon of metaphorical representation in language.

Representationalism

Generally, most of the issues presented in texts relate to the idea of representationalism. And even though this notion and its theoretical grounds are rarely mentioned in the texts, all the problems tackled can be placed within the grounds of representationalism.

Representationalism has been and maybe still is the most important paradigm in the research into mind and cognitive processes in both philosophy and cognitive science. However, it assumes that cognition is of representational character, the phenomenon of consciousness is the chief obstacle to representationalism since it is difficult to explain consciousness in representational terms (save the state of self-consciousness).

On the grounds of scientific research, and not only philosophical inquiry, representation still is not a widely accepted view on what it means for a cognitive system to represent something. However, "[T]he lack of a theoretical foundation and definition of the notion has not hindered actual research" (Vilarroya, 2017, p. 1).

REFERENCES

Daddesio, Th. C. (1994). On Minds and Symbols: The Relevance of Cognitive Science for Semiotics. Berlin: Mouton de Gruyter.

Jacobson, A. J. (2003). Mental Representations: What Philosophy Leaves Out and Neuroscience Puts In. *Philosophical Psychology*, *16*(2), 189–203.

Konderak, P. (2013). Mind, Cognition, Semiosis: Ways to Cognitive Semiotics, Lublin: Maria Curie-Skłodowska University Press.

Vilarroya, O. (2017). Neural Representation. A Survey-Based Analysis of the Notion. *Frontiers of the Psychology*, *8*(1458). doi:10.3389/fpsyg.2017.01458

Zlatev, J. (2012). Cognitive Semiotics: An Emerging Field for the Transdisciplinary Study of Meaning. *The Public Journal of Semiotics*, *4*(1), 2–24.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXV, nr 1 (2021), s. 9–34
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxv1.02

Article

JAKUB JONKISZ [*]

CONSCIOUSNESS, SUBJECTIVITY, AND GRADEDNESS[1,2]

SUMMARY: The article suggests answers to the questions of how we can arrive at an unambiguous characterization of consciousness, whether conscious states are coextensive with subjective ones, and whether consciousness can be graded and multidimensional at the same time. As regards the first, it is argued that a general characterization of consciousness should be based on its four dimensions: i.e., the phenomenological, semantic, physiological and functional ones. With respect to the second, it is argued that all informational states of a given organism are subjective (as they are biologically individuated), but not all are necessarily conscious. Finally, where the third question is concerned, in each of the four dimensions of consciousness a graded element is identified: quality of information in the phenomenological one, abstractness in the semantic one, complexity in the physiological one, and usefulness in the functional one. The article also considers certain consequences of the solutions proposed, as well as some practical applications of the 4D-view of consciousness.

KEYWORDS: graded consciousness, individuated information, subjectivity, dimensions of consciousness.

1. Introduction

Contemporary consciousness studies is a field that presents us with a multiplicity of more or less fundamental problems of both an empirical and a theoreti-

[*] Jagiellonian University, Institute of Psychology, Consciousness Lab. E-mail: kjonkisz@wp.pl. ORCID: 0000-0001-7221-4233.

[1] This article was supported by an OPUS grant from the National Science Centre of Poland (2017/25/B/HS1/01591).

[2] The article shares its main theses with the text published in Polish in *Filozofia Nauki* (Jonkisz, 2019).

cal kind (Dehaene et al., 2017; van Gulick, 2018). Of these, the most basic concerns the lack of an unambiguous characterization of consciousness itself. There is still no universally accepted description of the phenomenon of consciousness, or general definition of it, while the operationalizations employed in particular research cases often differ significantly (Jonkisz, 2012; Pareira, Ricke, 2009; Velmans, 2009). Consciousness may be said, at the very least, to be a concept lacking in sharply defined boundaries (in that its scope has not been clearly defined to date) or an ambiguous phenomenon. (Alongside this, the possibility also persists that it could potentially refer to multiple quite distinct phenomena; see Block, 1995; Irvine, 2012; Torrance, 2009). A closely linked question concerns the relationship between consciousness and subjectivity: are the conscious states of a given organism or system coextensive with its subjective states? Such an assumption, though by no means self-evidently valid, seem to be operative in many influential conceptions and theories of consciousness today (e.g., Block, 1995; Chalmers, 1996; Searle, 1992; 2000). Another currently important issue concerns the gradability of consciousness: i.e., the question of whether consciousness emerges in steps, or with an increasing intensity/sharpness, or rather appears suddenly in an all-or-none fashion (Andersen et al., 2016; Overgaard et al., 2006; Sergent, Dehaene, 2004; Windey et al., 2013; 2014). This problem is particularly interesting, given the multi-dimensional nature of consciousness, as certain researchers insist that it is very difficult to justify ascribing such gradedness to consciousness in respect of its manifold dimensions (Bayne at al., 2016).

The paper proposes certain solutions to the three problems just mentioned: namely, that of how to give an unambiguous characterization of the phenomenon itself, that of the relationship between consciousness and subjectivity, and that of the gradability of consciousness. The aim of this article is to present and justify those solutions in a condensed form (for more details, see Jonkisz, 2015; 2016; Jonkisz et al., 2017), while at the same time pointing out their consequences and related issues worthy of further study.

2. Characterizing Consciousness

Above and beyond the operational definitions used in specific research cases, or in the common-sense description of consciousness as a state of wakefulness contrasted with deep sleep, coma or state of general anaesthesia (e.g., Damasio, 1999; Searle, 2000), formulating an unambiguous, universal characteristic of the phenomenon itself remains a difficult and problematic matter (Torrance; 2009, Velmans, 2009). As a consequence, it is possible to identify as many as several dozen different meanings for the term, together with the corresponding kinds, types or varieties of consciousness being posited, within consciousness studies (Brook, 2008; Jonkisz, 2012; Pareira, Ricke, 2009). However, amidst all this diversity, which can sometimes lead to the conclusion that we are dealing not with one but with many different phenomena (Block, 1995, p. 227), it is possible to point to a relatively small number of recurring descriptive elements. Consciousness is quite often identified with a totality of subjectively experienced states or qualities (e.g., Block, 1995; Chalmers, 1995; Kriegel, 2006; Tononi, 2004). It is also claimed that consciousness consists in intentional—in the sense of being always about something (Searle, 1992; 2000)—first- or higher-order states (e.g., thoughts about thoughts, or perceptions of perceptions; see, e.g., Carruthers, 2016; Gennaro, 2005; Lycan, 1996; Rosenthal, 1986). Consciousness is also presented as a state generated by the specific brain processes—namely, widespread or more localized recurrent neuronal activity in the thalamocortical regions (e.g., Crick, Koch, 2003; Dehaene, Changeux, 2011; Edelman, 2003; Lamme, 2006). Finally, it has been characterized in terms of being a certain adaptation that allows its possessor to, among other things, more effectively adapt to new stimuli, solve problems, decide, understand and empathize (e.g., Baars, 2002; 2012; Cohen, Dennett, 2011; Damasio, 1999; Feinberg, Mallatt, 2013; 2016; Merker, 2005; Morsella, 2005). The aforementioned descriptions relate, in principle, to four different aspects or dimensions of consciousness (Jonkisz, 2012; 2015; Jonkisz et al., 2017). Firstly, they concern its phenomenological dimension, which includes the qualitative aspect of conscious states fully accessible only from the first-personal, or subjective perspective. Secondly, they relate to its semantic dimension, since the intentionality of both first- and higher-order states can be reduced to the semantic property known as referentiality (with different orders of reference; see Jonkisz, 2012; 2015; Pierre, 2003). Thirdly, they pertain to its physiological dimension, which concerns the mechanisms most likely to produce consciousness in a given organism. (In this respect, scientists like to point to specific neuronal correlates—so-called NCCs; see Metzinger, 2000). Fourthly, they concern to its functional dimension, which deals with the adaptive role of consciousness in a given organism's actions.

The vast majority of meanings, kinds, types or varieties attributed to consciousness either directly reflect one of the dimensions just mentioned or involve some sort of combination of them. Such a four-dimensional approach thus enables one to organize the different versions of the concept of consciousness into a clear taxonomy (Jonkisz, 2012; 2015). Apart from its elucidatory advantages, such an approach also has explanatory value, in that it allows us to define four distinct research problems. In relation to consciousness, one can ask: "Why and how is there something it is like to have it?" (the focus of explanation then being its qualitative characteristics, accessible from a subjective perspective); "How does it refer to anything?" (explanations thus concentrating on its semantic properties); "How does it emerge in a given organism or system?" (the aim being to understand its mechanism(s) of production); and, finally, "Why is it that these and no other states are conscious?" (the research focusing mainly on such states' pragmatic function).

If the physiological, phenomenological, semantic and functional dimensions actually exhaust the concept of consciousness as it is known to science, then they would seem to represent a reasonable starting point for an attempt at formulating a general characterization of consciousness. Pursuing this direction, consciousness may be broadly described as a *state co-occurring or caused by specific neurophysiological processes, in which a given organism experiences certain (referential) contents related to its actions.*[3] However, even if correct, such a description is surely too general, in that it, too, allows for multiple different interpretations, where these ultimately may bear on the actual scope of applicability of the concept. Indeed, such a situation is observable at the moment in consciousness studies, where one can encounter both very narrow construals of consciousness (e.g., Carruthers, 1998; 2018—an author who holds that it is in principle only possessed by humans) and quite broad ones (Feinberg, Mallatt, 2013; 2016—authors who claim that forms of consciousness are possessed by evolutionarily very old organisms, such as, e.g., sea lampreys or insects). On an ultra-wide construal, consciousness is attributed not only to the majority of living organisms, but also to certain artificial systems, sometimes even very simple ones (e.g., Tononi, 2004; 2008; 2010—along with his integrated information theory or IIT, in which even light-sensitive diodes are claimed to have something more than a zero degree of consciousness). It therefore seems quite important to furnish grounds for accepting some reasonable limitations: limitations that may help to avoid such radical shifts in the range of what may count as instances of consciousness in the context of contemporary conceptions (Jonkisz, 2015).

Such initially imposed limitations may consist in a determination of the minimal requirements for potentially conscious organisms or systems (these are referred to as “global limitations”; see Jonkisz, 2015). In this regard, my research yielded the hypothesis that only those organisms or systems that individuate information are capable of producing subjective perspectives. This is based on two assumptions. The first is that information, as a category, is superordinate to (i.e., encompasses) consciousness, in that all states of consciousness are informational states, but not all informational states are conscious. Such an idea is by no means a one-off in contemporary research: for example, Koch and Tononi (2013) present a similar view, according to which consciousness consists in integrated informational states, as does Earl (2014), who claims that all states of consciousness are nothing more than various forms of information. The second is that, since consciousness is a phenomenon so far observed only in nature, the concept of information should also be naturalized—meaning that it will then be interpreted as a state of an organism that carries biologically justified value/meaning for that organism.[4] This assumption eventually led me to the conclusion that all information possessed by a given organism must be unique, because its form, meaning and pragmatic functions have been shaped (and continue to be

[3] Such a description presents the phenomenon of consciousness as being limited to living creatures; however, if the term “organism” is replaced with “system”, and “neurophysiological process” with “mechanisms”, it no longer excludes the possibility of artificial or machine consciousness (Hollande, 2003; Torrace et al., 2007).

[4] This interpretation is intended as provisional, in that it awaits further justification in separate studies.

so) by a coinciding of multiple evolutionary, developmental and environmental factors that will be distinctive for just that very creature itself. On this account, it seems that all informational states available to a given organism must have undergone a complex individuation process leading to the creation of its unique, private perspective—a process that is most probably a necessary condition for subjectivity. (This thread will be further elaborated in the next part of the present text).

Further limitations may consist in a determination of the conditions that must be met for something to count as an occurrence of consciousness in a given organism or system (local limitations). The real aim here is to find a certain contrast for consciousness: i.e., a significant difference between conscious and unconscious information processing. Such a search is indeed underway, as researchers try to spot the difference in each of the dimensions described above. As regards the physiological dimension, what we are looking for is a contrasting neuronal mechanism or activity pattern. However, despite many important discoveries, no consensus has been reached on this, either regarding an exact location or with regard to a specific, consciousness-providing activity (Hohwy, 2009; Metzinger, 2000; Noë, Thompson, 2004). The contrast between conscious and unconscious states is also not very clear in the semantic dimension. This is partly because all informational states, even unconscious ones, must refer to or mean something for the organism or system in order to count as informative at all. Admittedly, some scholars point out that the form of such a state (e.g., whether it is a thought, perception or representation) or its order of reference (first or higher-order) may be decisive (Carruthers, 2016; Kriegel, 2007), but again there is no consensus here. Neither does it seem that the phenomenological dimension brings into play any conclusive difference. Most obviously, this is because it is difficult to verify whether the organism is subjectively experiencing anything at all at a given moment. (We must rely here on behavioral measures that mostly measure reports not occurring simultaneously with the experience and so potentially influenced by many different processes; see Timmermans, Cleeremans, 2015). Yet this is also because even subjective experience itself does not have a clearly identifiable set of characteristics: if we do not know exactly what it is, then we cannot be certain whether someone other than ourselves has it or not. Meanwhile, such notions as “qualia” or “phenomenological consciousness”, which are often invoked at this point, only worsen the situation (Block, 1995; Dennett, 1988,). Furthermore, it is even possible to find grounds for arguing that unconscious states are also subjective. (See the next part of this text, below, as well as Jonkisz 2009; 2016; Neisser, 2006; 2015). Partly because of these problems, the functional dimension is seen as being, at least for now, the more rational option when it comes to searching for a contrast between unconscious and conscious information processing. There is not enough room here to analyse the various conceptions and disputes that surround the function of consciousness itself. (This issue will, though, be addressed in a little more detail in the third part here; see also Baars, 2002; 2012; Cohen, Dennett, 2011; Hesselman, Moors, 2015; Morsella, 2005; Merker, 2005). Nevertheless, the assumption that conscious information processing represents an evolutionarily valuable adaptation seems self-evidently reasonable—after all, organisms must have been more efficacious and statistically more successful when acting consciously, otherwise the ability to do so probably would not have survived (Feinberg, Mallat, 2013; 2016; Griffin, 2001; Hassin, 2013; Lindahl, 1997). On the basis of just this relatively straightforward assumption, to the effect that consciousness yields a certain advantage in respect of a given organism's actions, and without specifying what the actual function in question is, the following hypothesis then seems acceptable: out of all of the informational states accessible to a given organism or system, the states that reach consciousness will most likely be those that are functionally the most useful in action at a given moment, from the subjective point of view of that organism or system (Jonkisz, 2015; 2016; Jonkisz et al., 2017).

Ultimately, on the basis of hypotheses limiting consciousness globally (to systems that individuate information) and locally (to informational states most useful in action), consciousness may be characterized very concisely as "individuated information in action" (Jonkisz, 2015; 2016). Some of the consequences of this characterization will be discussed in the closing section below, while in the next part I shall offer a slightly more precise discussion of the concepts of individuation and information.

3. Consciousness and Subjectivity

In order to know what-it-is-like to see a red rose, smell its scent, or feel the prick of its spike, one must consciously experience the sensations oneself. It is generally assumed that any such qualitative character of consciousness is only available from the internal or private perspective of the subject: i.e., only subjectively. From an external perspective, or objectively, we can observe certain forms of accompanying behavior and physiological parameters correlated with these experiences, although in the case of humans we also encounter relations to verbal utterance. (It is worth noting that in this context the "subjective versus objective" distinction specifically refers to the form of cognitive accessibility involved; hence, it may be said to be understood epistemically).

Many researchers consider the subjectivity of consciousness, i.e., its phenomenological dimension, a particularly "hard problem", treating explanatory problems related to other dimensions as relatively easy to solve by comparison. It has even been argued that because science is unable to fully answer the question of "what-it-is-like to experience something", when it comes to qualia or so-called phenomenological consciousness we are basically faced with what is known as an "explanatory gap" (Bayne, 2009; Block, 1995; Chalmers, 1995; Dennett, 1988; Jackson, 1982; Kriegel, 2006; Levine, 1983; 2001). It is hard not to agree here with Edelman, who states that the hard problem, put this way, "does not require a solution, but rather, a cure" (Edelman et al., 2011, p. 5). The assumption that science should actually encompass the subjectivity of conscious experiences, furnishing complete knowledge about what-it-is-like to have them, is a category error: this kind of knowledge is accessible only from within an experiencing system, not from any scientific statements or theories (Pigliucci, 2013). Scientific descriptions of subjective experiences, even of the most detailed kind, will not generate these experiences—that much is obvious. However, this need not mean that science is unable to explain subjectivity (Baars 1996, p. 2011; Edelman, Tononi, 2000, pp. 139–140). So how should science, which is essentially objective, seek to explain subjective consciousness? Besides, where possible, an unambiguous and precise determination of the concepts and research objectives involved, we usually expect from science explanations of either a functional or a mechanistic nature, or both. In this instance there is no reason to expect otherwise, so the actual goal must be to understand the functions and mechanisms of subjective consciousness. Moreover, since the research in question aims to shed light on a natural phenomenon (in that consciousness occurs in nature), both of these aspects should be interpreted naturalistically. Therefore, in asking about functions, we should be looking to identify a possible adaptive role for subjective awareness, in the sense of any advantages it might provide in the context of action of an organism. On the other hand, when it comes to mechanisms, one may here ask two questions—one posed at the evolutionary level, the other from a physiological perspective. The first would be this: When and how did subjective consciousness develop amongst living organisms? (Feinberg, Mallat, 2013; 2016.) The second, on the other hand, would be the following: What processes are responsible for the production of consciousness in a given organism? (Bisenius et al., 2015; Edelman, Seth, 2009; Koch et al., 2016) Below, I shall put forward hypotheses pertaining to both the functions and the mechanisms leading to the formation of subjectivity.

As was mentioned in the previous section, information is regarded as being superordinate (in conceptual-hierarchical terms) to consciousness, in that all conscious states are informational states, but not vice versa. That commitment receives a brief justification below. As a starting point, I shall accept, at least in broad terms, a characterization of informational states based on information integration theory: an informational state is a state of a system differentiated by that system from its other states, where a state of a system is determined by the interaction of its elements (Koch, Tononi, 2013; Tononi, Koch, 2014). Described thus, informational states must necessarily include all states of consciousness, as in order for a given organism-system to become conscious of something, it must somehow identify the latter, or at least differentiate it from other things (the system must detect the signal or stimuli and integrate it/them as a new whole). At the same time, many studies suggest that much of the information processed by our nervous systems is not conscious, and this applies not only to "lower-level" but also "higher-level" information processing, such as engages the prefrontal areas of the brain usually associated with fully formed consciousness (van Gaal et al., 2012). It is also argued, that even executive, top-down control of behavior (Kiefer, 2012) and fully integrated states (Mudrik et al., 2011; 2014) might be carried out unconsciously. In the context of the characterization of informational states just offered, it can be said that not all states differentiated by a given organism-system become conscious for that system.[5] Ultimately, all the states of consciousness of a given system form a subset of the informational states available within that system. Yet if there are both conscious and unconscious informational states available within a certain organism or system, what is their relationship with subjectivity? Could only conscious states be subjective, as they surely are, or is it perhaps the case that all informational states of a given system possess this feature? The answer will largely depend on the notion of subjectivity applied—in this context, one characterized as "availability limited to the internal perspective of a given organism-system". Already, in (Jonkisz, 2009), I argued that the formation of subjectivity, understood this way, can be explained by pointing to the structural and functional uniqueness of organisms. In my recent studies (Jonkisz, 2015; 2016), as was already mentioned above, the concept of individuation has emerged as crucial, so I should now describe this in more detail.

Generally speaking, individuation is understood here as a complex selection process that includes both sources of information and the informational states themselves. In principle, it can be said that it begins at the evolutionary level, because the availability of information of a certain type is determined by the morphological and physiological equipment of a given species. Each and every creature is, quite simply, limited: for example, by its sheer manner of getting around (so flying, say, will furnish different informational possibilities than swimming or walking), but also by the type, amount and sensitivity of its receptors (which, for instance, only allow for the detection of specific wavelengths and frequencies of light and sound, or specific chemical compounds). As a consequence, organisms are able to detect just certain kinds of stimuli and process information of only a certain type (i.e., those which proved most efficacious for their biological ancestors—if we may be permitted to thus simplify the logic of evolutionary justification). The process of individuation continues as information, reduced at a phylogenetic level to specific resources, is subjected to further specifications, being modified by epigenetic factors (inherited by subsequent generations), changed by certain social components (e.g., different values and meanings within specific groups of organisms), and also reflecting specific environmental conditions, in force at a given moment in time (Ballestar, 2010; Bossdorf et al., 2008; Fraga, 2005; Migicovsky, Kovalchuk, 2011; Swaddle et al., 2005). Consequently, information acquires more and more specific forms and meanings, becoming virtually unique at the level of a given phenotype. Moreover, the differentiation of specific informational states, experienced thus and in no other way by given organism, ultimately depends on multiple individual factors, such as the following: the current state of the organism (e.g., biochemical parameters of its

[5] In practice, this may for example mean that not all new activity patterns, even integrated within the cortical regions, inevitably result in conscious experience—as is indeed often the case (Kiefer, 2012; Mudrik et al., 2011; 2014; van Gaal et al., 2012).

nervous system), its being located in some specific surroundings (i.e., limitations pertaining to the availability of space, time, relationships, engagements, etc.), its individual history (given that already experienced states will influence future states), and its currently extant decisions, challenges, plans undertaken, etc.[6] All these will be reflected in the constantly changing structure of its body—in particular, in the network of connections and activity patterns in the nervous system. That is why there are, in fact, no two identical nervous systems: even the brains of identical twins differ significantly (Freund et al., 2013; Frith, 2011; Marti et al., 2011; Pfefferbaum et al., 2004; Valizadeh et al., 2018). Once again, we are led here to agree with Edelman, who states that, as a result, "[a]t any given moment, a process of integration of collective neuronal activity generates an interwoven pattern of responses unique to a particular animal at that particular moment of time" (Edelman et al., 2011, p. 3).

In conclusion, we may assert that as a result of such a complex and extended process of individuation involving multiple levels—be they phylogenetic or ontogenetic, genetic or epigenetic—biological systems are structurally and functionally unique, and therefore operate in highly individualized informational spaces. Hence, any informational state that a given organism is capable of differentiating will in fact be available only at a given moment, only for that particular system, and only from its own unique and, in effect, private cognitive perspective. Ultimately, if subjectivity is understood in terms of availability limited to the internal perspective of a given organism-system, we may conclude that all of the informational states of a given system are subjective, regardless of whether they are conscious or not.[7] Consequently, subjectivity cannot be taken to be a feature specific only to states of consciousness, as its range of instances turns

[6] Many studies have confirmed the importance of both the top-down and bottom-up effects of bodily factors on information processing (e.g., Fleming et al., 2010; Pfeifer et al., 2014; Rochat, 2011; Shimono et al., 2012; Theeuwes, 2010; Zhou et al., 2013).

[7] As a reviewer has rightly pointed out, this talk of "availability limited to the internal perspective" needs to be fleshed out in more detail, since it bears the weight of important conclusions drawn in the present article. To be as concise as possible: the subjective character of conscious states should not be understood coextensively with their phenomenal character (the fact that they are experienced), as otherwise statements such as "consciousness is subjective because it is accessed/available only as experienced (only from the first-person perspective, from the inside, from within, etc.)" will be circular. Hence, a given state's being subjective (i.e., internally available for a given organism/system) cannot be coextensive with its being experienced. But in that case, what will it mean for an informational state to be subjective, yet not conscious? A given system's being informational will be understood here along Tononian lines: i.e., as "differentiated by that system" or "detected by that system". (A Shannonian take on this, involving uncertainty reduction in noisy-channels, will also be applicable here.) Ultimately, to be an informational state that is subjective but not conscious will mean that apart from being differentiated, it also has to be "available only from within a given system" (where this is explicated here in terms of "information individuation") but not experienced (i.e., with no phenomenology presenting itself).

out to be wider.[8] At the same time, such a process of individuation can be construed as a hypothetically posited natural mechanism, responsible for the development of subjectivity in the animal world. But, of course, this will not then serve to explain the emergence of consciousness.

It will quite likely prove possible to discern not only mechanistic differences between consciousness and subjectivity, but also functional ones. First, however, we should address the question of what adaptive advantage such a highly individualized perspective may provide for an organism. By way of justification, we may appeal in our answer to the rather obvious assumption that the selection of effective ways of action, combined with their rapid adaptation to the changing conditions of the moment, represent key adaptations for any organism. It is also quite plain that, in a complex and ever-changing environment such as we are dealing with here, organisms are potentially capable of distinguishing an infinite amount of information in an infinite number of states. In connection with such an "informational overflow", and the need for effective, but also swiftly executable actions, we may posit the existence of an evolutionary pressure to filter out the least valuable sources of information and choose the most useful ones from those available. This scenario is extremely well suited to the very process of individuation just described—one which, through a complex selection of possible informational states, leads to the emergence of subjective perspectives. I would thus assert that subjectivity is, most likely, an adaptive response to informational overflow, with its basic function being the selection of information that is the most valuable from the perspective of a given subject-organism-system (Jonkisz, 2016). As regards the function of consciousness, as was already indicated (in the previous section), this is taken to be manifested in action. (Its function will be described in more detail in the next section).

4. The Gradability of Consciousness

The issue of the graded versus the dichotomous nature of consciousness hinges on multiple heterogeneous factors, and presents itself as being even more complex than the issues discussed above. The very concept of consciousness utilized in the relevant research can play a determining role in this regard. For example, in so-called Higher-Order Theories or HOTs, one may incline towards treating consciousness as a property that appears suddenly, because arriving at a higher-order state by a given subject (be it perception of perception, representation of representation, or thought about thoughts) is something that takes place all at once rather than gradually (Carruthers, 2016; Gennaro, 2005; Lau, Rosenthal, 2011; Lycan, 1996; Rosenthal, 1986). On the other hand, gradability seems quite natural as something to embrace in approaches associated with the so-called "integration consensus" (Seth, 2009; Tononi, Koch, 2014) or using non-

[8] Such a conclusion regarding the existence of "unconscious subjectivity" may strike one as surprising, but is not isolated (Farisco, Evers, 2017; Neisser, 2006; 2015).

report methodologies (Tsuchiya et al., 2015). These conceptions utilize numerical measures of consciousness, and also quite often take non-human creatures and even artificial systems to be capable of being conscious. The tendencies towards viewing consciousness as graded or dichotomous may also be at least partly subject to polarizing influences stemming from the particular field of research being brought to bear on this topic. For example, in psychiatry or neuropathology, the notion of levels of consciousness is quite widely accepted (it being depicted in different scales of consciousness; see Giacino, 2005; Schnakers et al., 2008; Teasdale, Jennett, 1974), whereas in contemporary philosophical and psychological approaches this is by no means obviously the case (Bayne et al., 2016). Finally, the methodology used in research may exert an influence on the answer given: for example, within the so-called "subjective measures of awareness" (e.g., Timmermans, Cleeremans, 2015), what may be at least partially responsible for the outcomes is the simple choice of scale used in the experiment, or the mere selection of specific tasks or stimuli presented to participants (whether, for example, those are more or less complex, more or less abstract, induce higher or lower processing levels, etc.).[9]

It is quite likely that the difficulties involved in answering the question about the graded or all-or-none nature of consciousness are also caused by the fact that it is simply not clear what we are asking about. In other words, it is not being specified precisely enough what it is that can actually appear suddenly or emerge in steps or with varying intensity—whether we mean by this the subjective content of consciousness, or some specific physiological parameters of that state, or something else. The matter becomes even more complicated if we take on board the assumption that consciousness is a phenomenon having (at least) four different dimensions: i.e., phenomenological, physiological, semantic and functional (as described in the first section here). It is not clear whether gradedness should be visible in each of the dimensions independently, or rather in all of them at the same time—so, should we be looking for four different hierarchies of levels of consciousness, or rather for just one, somehow averaged out across these? In any case, considering the multidimensionality of consciousness, one must accept that individuals cannot be "ordered on the basis of how conscious they are, just as they can be ordered on the basis of their age, height, or blood pressure" (Bayne et al., 2016, p. 406). Even so, is it really necessary to draw critical conclusions from this line of thinking for all of the graded approaches, as the authors of the text just quoted do? Below, I shall put forward some practical guidelines for how to reconcile the four-dimensional conception of consciousness with gradability: more specifically, I will show what is or could be

[9] For example, it has been shown that stimuli/tasks that manifest higher levels of processing (like semantic discriminations) will more often result in the subject's believing the emergence of conscious experiences to be something occurring on an all-or-none basis, whereas low-level features (e.g., shapes, locations) result in experiences that are taken to appear gradually (Windey et al., 2013).

graded in each of the dimensions and how to measure it (in the sense of describing possible or actually existing methods of measurement; see Jonkisz, et al., 2017).

What, then, can be pointed to as being a graded element in the phenomenological dimension? Just to recall, this dimension refers to the qualitative characteristics of states of consciousness, which are accessible only from the private perspective of the subject (subjectively). During conscious seeing, hearing, smelling, bodily sensation, thinking, imagining, etc., we experience different objects, sounds, colors, feels, smells, etc. Actually, these experiences appear as more or less vivid, sharp, intense, clear, rich, detailed, etc. It can therefore be concluded that, if something is graded here—i.e., it decreases or increases—then we may point to these very qualities: i.e., vividness, sharpness, intensity, etc. Generalizing this idea, it can be assumed that the quality of experienced states of consciousness is the gradable element in the phenomenological dimension, in the sense that the states possessing a higher quality grade would present themselves as being more vivid, sharp, intense, etc., while those with low quality would show up as less clear, unclear, blurred, barely perceptible, etc. The idea seems quite plausible, but is it possible to actually measure the phenomenal quality garade? In fact, a variety of methods already exist in consciousness studies that are applied to measure this parameter. These include objective methods, based on behavioral criteria, signal detection data and/or neuronal activity patterns analysis (Heavey, Hurlburt, 2008; Tsuchiya et al., 2015), and subjective methods, based on the analysis of reports concerning one's own conscious experience, as given by participants (Overgaard, 2015; Overgaard, Sandberg, 2012; Wierzchoń et al., 2012). For example, the results obtained using so-called subjective measures of consciousness (Wierzchoń et al., 2014) suggest that consciousness is in fact subjectively graded in certain cases—to be more specific, in these cases participants use all available grades of a given scale to report on the quality of experience they have had, following a specific presentation of stimuli on a screen (Overgaard et al., 2006; 2010). Nevertheless, in other studies employing similar methodology, conscious experience seems to be dichotomous, with respondents in such cases invoking extreme ends of the scale and indicating that the presented stimulus was either clearly visible or that there was no experience of it whatsoever (Sergent, Dehane, 2004). Therefore, some researchers claim that from a subjective perspective consciousness can actually be both: i.e., sometimes graded and other times dichotomous. (It is assumed that the hypothesis of the level of processing may help to explain this claim; see Windey et al., 2013; 2014).

Apart from its subjective characterization, consciousness also stands in an objective relation to what it refers to or is about (Legrand, 2007, p. 577). This referentiality of consciousness forms the basis of its semantic dimension. So, is there any chance for a graded element to show up here, too? Conscious states may refer to anything that we sense, feel, think of, remember, imagine, etc. Yet the reference may be either more direct, as in cases where one is just conscious of the sheer sensations, feelings, thoughts, etc., or more abstract, as in cases where one is also conscious of the sensing, feeling, thinking, etc., too. Hence, it can make sense to say either that the subject *X* is conscious of *Y*, or that the subject *X* is aware of being conscious of *Y* (Jonkisz, 2012; 2015; 2016, where five consecutive orders of consciousness are described). In dealing with the first sort of instance, researchers apply, among others, such terms as "first-order consciousness" or "non-reflective consciousness", whereas in the second case they talk of such things as "higher-order consciousness", "reflective consciousness", "introspective consciousness", or "metacognition" (Armstrong, 1979; Lau, Rosenthal, 2011; Morin, 2006; Overgaard, Sandberg, 2012). Ultimately, on the basis of the order of reference involved, this relation may be considered more or less abstract, so abstractness can be considered a graded element within the semantic dimension. Although higher-order states and metacognition have been studied empirically (e.g., Fleming, Lau, 2014; Middlebrooks, Sommer, 2012), there are no measures of abstractness itself in use as of today. However, one could expect that performing certain types of activity, or completing certain kinds of task, would result in the occurrence of more or less abstract states. For example, in procedures that apply subjective measures of consciousness, a visual stimulus is displayed to a participant in near-threshold time durations (e.g., simple geometrical shapes, strings of letters, numbers, or more complex objects like male or female faces). The participant is subsequently given an identification task, followed by the task of assessing the experienced quality of the image (perceptual awareness scale or PAS) or rating their level of confidence in what they have just seen (confidence ratings or CR) on a scale of four grades (Dienes et al., 1995; Ramsøy, Overgaard, 2004; Wierzchoń et al., 2014). One may conclude that first-order visual consciousness is not sufficient to complete these tasks, as in order to assess (first-order) visual experience one needs to be aware not only of the visual object itself, but also of the experienced quality of the (higher-order) seeing of that object.[10] In practice, it may prove useful to create the sort of procedures that will allow us to assess more precisely the order of reference invoked by a given task.

Gradability in the physiological dimension seems to be a quite obvious affair. Thanks to the development of new research methods combining neuroimaging with electroencephalography (EEG) and transcranial magnetic stimulation (TMS), we find ourselves increasingly well placed not only to determine, but also to understand, the brain mechanisms and distinctive neuronal activity patterns associated with occurrence of consciousness (Bandettini, 2009; Bisenius et al., 2015). Analysis of these patterns enables one to assess the level of integration of the various brain regions cooperating at a given moment (mostly on the basis of the synchronization of the activities involved), as well as the range of differentiation of these regions (in the sense of assessing their heterogeneity across different portions of the cortex). The relationship between integration and differentiation is currently being intensively studied, and at least three different

[10] Thus, it remains a matter of dispute whether the quality of conscious experience, or in fact the quality of metacognition, is what is actually being measured by means of PAS and CR (Wierzchoń et al., 2014).

ways of enabling its numerical determination have been proposed. In integrated information theory, this dependence is reflected by the Φ-value, with the theory working on the assumption that the higher the Φ-number is, the greater will be the ability of a given system to integrate information (Tononi, 2004; 2008; 2010; Tononi et al., 2016). Anil Seth (2008), meanwhile, proposes the so-called "causal density" value or "cd", calculated on the basis of the analysis of interaction between elements of the neuronal network relevant at a given moment (Barrett, Seth, 2011). Lastly, Missimini and his colleagues have developed the so-called Perturbational Complexity Index (PCI), in which the cortical response to intentional perturbations evoked by Transcranial Magnetic Stimulation (TMS) impulses is assessed (its EEG complexity pattern being calculated using Lempel-Ziv; see Casali et al., 2013). Simplifying the overall notion underpinning such models, we can state that according to such proposals the more complex the activity patterns (indicated in a higher Φ, "cd" or PCI value), the higher the probability of the occurrence of consciousness—or the higher its level. Ultimately, it can be assumed that the physiological gradability of consciousness is reflected in the overall complexity of the activity patterns involved. However, one should keep in mind that these methods assume a correlation with consciousness: in other words, while not very likely, one might still obtain a high numerical value unaccompanied by consciousness.

On a four-dimensional approach, one can certainly still inquire into functional gradability. So is it possible, in this dimension, to point to some parameter or other that increases and decreases, or appears in steps? The issue is not straightforward, because there is no consensus even as to specific function supposedly performed by consciousness (Hesselman, Moors, 2015). However, as was already mentioned, the assumption that consciousness is an evolutionarily valuable adaptation seems quite obvious (Feinberg, Mallat, 2013; 2016; Griffin, 2001; Hassin, 2013; Lindahl, 1997). The value of adaptations is reflected in the abilities they provide for organisms, so we can ask what it is that the conscious processing of information actually furnishes. There have been many proposals regarding this matter: for example, that consciousness enables learning, decision making, action planning, problem solving, etc. However, it has been argued that all such functions could also be performed in the absence of consciousness (e.g., Hesselman, Moors, 2015). Recently, though, one idea does seem to have gained fairly wide acceptance in the context of the currently predominating theoretical approaches: it is that conscious processing enables the integrating of signals and information from various systems (e.g., sensory, motor, memory) and different cortical regions (Baars, 1994; 2002; Baars et al., 2013; Dehaene, Changeux, 2011; Dehaene et al., 1998; Dehaene, Naccache, 2001; Edelman, 2003; Edelman, Tononi, 2000; Edelman et al., 2011; Tononi, 2004; 2008; 2010; Tononi, Koch, 2014; Seth et al., 2005; Seth, 2009; Palmer, Ramsey, 2012). Although unconscious processing is much faster and more economical (in terms of energy consumption), information integration pays off, as it provides significant flexibility in action: i.e., it enables ongoing adaptation of behavior to changing external situations and internal preferences (Baars et al., 2013; Pally, 2005; Seth, 2009). It seems, then, that right now flexibility may be considered the most plausible function of consciousness. The usefulness of conscious processing would then be directly proportional to the flexibility required in a given activity: i.e., it would increase when the demand for flexibility grows, and decrease when flexibility is not needed (as, for example, in repetitive actions). Ultimately, the varying usefulness of conscious processing seems to be a good candidate for a graded element in the functional dimension. At the present stage of my research, functional gradability, as determined by the degree of usefulness of conscious processing, can thus be entertained as a reasonable preliminary hypothesis. However, it should be pointed out that any potential measures of usefulness would have to take into account subjective factors (individual preferences, biases, aims, motivations, etc.) relating to previous experiences, as well as objective ones dictated by the actual state of the organism itself (available energy, possible behavioral responses, available sensory inputs and sensitivity, etc.) and by environmental conditions (available time, space, relations, interactions etc.).[11]

5. Consequences

Characterizing Consciousness. On the conception presented above, a key role is played by the differentiation of four dimensions of consciousness: i.e., phenomenological, semantic, physiological and functional. What we have found is that apart from its explicatory usefulness (enabling us to taxonomize the concept of consciousness), this set of distinctions also serves to bring to light important explanatory and methodological values (enabling to identify four important research problems). From this four-dimensional perspective, consciousness has been broadly characterized here as a state co-occurring with or caused by specific neurophysiological processes, in which a given organism experiences certain (referential) contents related to its actions. Moreover, global and local limitations imposed on such a conception have finally allowed us to characterize consciousness as individuated information in action.

As a consequence of the above, we must consider all biological and (even) artificial systems capable of utilizing individuated informational states in action to be (at least to some extent) conscious. Characterized thus, the real range of consciousness depends, however, on the way the terms involved are interpreted. It is obvious, then, that in order to achieve practical usefulness for such a concept, it will be necessary to introduce more precise guidance regarding the notions of

[11] Such assumptions seem to fit well with so-called Bayesian Brain models, in which cognitive systems are seen as a kind of inference-generating or predictive machine. In that context, conscious states could be described as those whose predicted usefulness ranks highest from the system's own perspective. The evaluation of usefulness could then be modeled using Bayesian statistics. Even so, how our nervous systems actually do this remains debatable, the principal issue being whether brains really quantify probability, or instead somehow test the expected efficacy of actions (Sanborn, Chater, 2016; Seth, Friston, 2016).

information, individuation and action. Otherwise, it will prove difficult to give a reasonable answer to the fairly obvious counterargument that, after all, not every instance of individuated information in action need be conscious. The issue requires separate, much more extensive research. However, it seems that we already now have reasons for thinking that the notion of information used in consciousness studies must be naturalized, as only this will allow us to properly register its biological uniqueness and/or individuatedness. In the light of these findings, we can now point to a direct connection with the function performed by conscious informational states in relation to a given organism's actions (namely, enabling flexibility), where this also seems to constitute a valuable achievement.

Consciousness and Subjectivity. In consciousness studies, the notion of subjectivity is typically understood in terms of privileged access, in the sense that qualitative characteristics of conscious experiences are taken to be accessible only from the first-person perspective of a given organism or system. The conception proposed here sheds light on the relationship between such subjectivity and consciousness. At first, it assumes that all states of consciousness are informational states, but not vice versa (i.e., not all informational states are conscious). It argues, then, that both the sources of information and the very informational states available to a given system undergo a complex process of individuation (with this process being justified functionally, as an adaptive response to the overflow of possible informational states). As a consequence, all informational states of a given system are individuated—which, de facto, means that they are accessible only from the perspective of this particular system: hence, they must be considered subjective (in the sense described above). Ultimately, this leads to a rather controversial conclusion about the existence of subjective but unconscious informational states. If the proposed line of argument is valid, and subjective states are not coextensive with conscious ones, then a characterization of states of consciousness in terms of their subjectivity or qualitative character (such as is quite common in contemporary conceptions) turns out to be inadequate. My proposal, on the other hand, also enables one to point to functional differences between consciousness and subjectivity. It has been argued that while the function of consciousness is manifested in action, in that it confers flexibility on the latter, the basic function of subjectivity should be considered to be the selection of information valuable from the perspective of a given organism. The process of individuation may, in addition, be considered an evolutionary mechanism leading to the emergence of subjective perspectives; yet that does not explain the emergence of consciousness. Ultimately, one can argue that the scientific—i.e., functional and mechanistic—explanations of subjectivity and consciousness simply differ. Despite the fact that all conscious states are indeed subjective, a straightforward identification of the so-called hard problem of consciousness with its subjective character or its phenomenological dimension now seems to fall short of being conclusively justified, to say the least.

The Gradedness of Consciousness. Gradational approaches enable more adequate descriptions of various forms of consciousness, including non-human consciousness, consciousness in the wake of injury to the brain, various neurological disorders, consciousness as exhibited at different developmental stages, and so on. On the other hand, consciousness is a very complex, multidimensional phenomenon—one that, as Bayne and his colleagues have rightly pointed out (2016), poses serious theoretical and practical problems. Those problems have been defined here in terms of two key questions: What is, or can be, graded with respect to consciousness, and can we measure it? In order to answer the first of these, the four-dimensional conception of consciousness was adopted, and then in each of the dimensions a graded element was identified—this being, respectively, quality of information in the phenomenological dimension, abstractness in the semantic dimension, complexity in the physiological dimension and usefulness in the functional one. As a consequence, consciously processed information has the potential to differ in at least four respects: with respect to its grade of quality, order of abstraction, level of complexity, and degree of usefulness. So far, two of these parameters have been measured in practice, these being the experienced grades of quality of conscious states (e.g., by means of report-based procedures, such as subjective measures of awareness) and the complexity levels of their neuronal underpinnings (e.g., by analyzing activity patterns with respect to their integration and differentiation in terms that enable us to calculate their Φ, “cd” or PCI values). There are, however, no practicable ways to measure orders of abstraction reached by conscious states, or degrees of usefulness achieved in the context of a given organism’s actions. Yet certain preliminary proposals and limitations regarding these issues have, I think it is fair to say, already been successfully marked out.

It is worth noting that four-dimensionally graded consciousness will not give rise to a linear scale, since a given organism or system may be ranked differently in each of the posited dimensions (e.g., simple visual images may be experienced with high-quality, while at the same time being not especially abstract semantically and exhibiting rather low levels of physiological complexity). Despite being seemingly complicated, such an approach is advantageous on many grounds: in explaining, predicting and putting forward hypotheses. For example, it is possible to more adequately describe consciousness in such states as *blindsight* or *locked-in syndrome*. As far as the first of these is concerned, it may be said that a blindsighted person would most likely have a very low or zero quality grade of visually experienced images. (Such persons usually claim that they do not experience any clear images; see Sahraie et al., 2010). In spite of lesions (usually located in the primary visual cortex or V1), visual information in the brains of such people is still processed in a way sufficient for them to guess what they see (with an above-random level of accuracy) and navigate efficiently in previously unknown spaces, avoiding obstacles. Hence, the degree of usefulness of such impoverished visual information is definitely not zero. It may also be plausibly argued that in blindsight the physiological complexity of the activity patterns involved can be pretty high, while visual information nevertheless only reaches the first-order level of abstraction (in that it only ever refers directly to perceived objects and lacks higher-order information about what is perceived or the very perception itself). On the other hand, when locked-in syndrome is considered the situation clearly looks different. A person who is in that condition may experience states with high quality (e.g., feeling a very sharp and localized pain, seeing and hearing clearly, etc.), and present neuronal activity patterns whose complexity does not deviate from the norm. Such a person could not only be conscious of different perceptual objects, but may also be aware of being conscious of them: consciousness as exhibited by locked-in patients does not seem any less abstract than in a normal state. However, patients that remain in this state are not able to practically perform any motor actions—except for vertical eye movements and, sometimes, blinking (Laureys et al., 2005; Schnakers et al., 2008; Smith, Delargy, 2005). Hence, the degree of usefulness of most of the information consciously available will be rather low for such a person, at least in sensorimotor-related terms. The examples described here can also be presented graphically (see Figure 1 for blindsight, and Figure 2 for locked-in syndrome).

A final closing question that I would like to raise here is this: If any of the four parameters is rated zero (e.g., quality in blindsight), should a state still qualify as conscious? Intuition would, I think, undoubtedly suggest a negative answer, the assumption being that a state is only conscious if it receives a non-zero result in each of the four dimensions. Yet properly justifying such a conclusion is hardly a straightforward matter, and constitutes yet another objective worth further study.

Figure 1

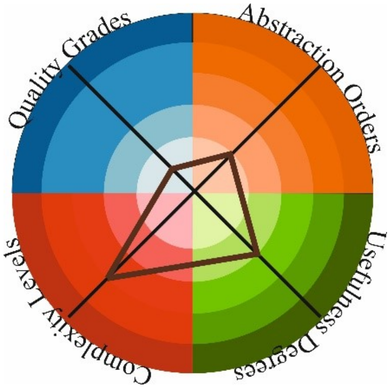
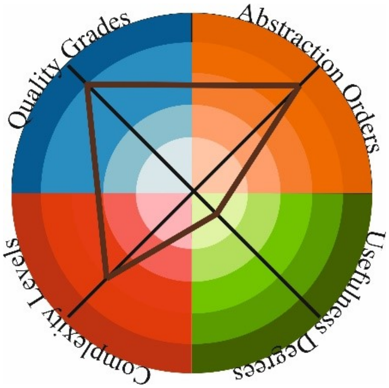


Figure 2



REFERENCES

Armstrong, D. M. (1979). Three Types of Consciousness. *Ciba Found Symp.*, *69*, 235–253.

Baars, B. (1996). Understanding Subjectivity: Global Workspace Theory and the Resurrection of the Observing Self. *Journal of Consciousness Studies*, *3*, 211–216.

Baars, B. (2002). The Conscious Access Hypothesis: Origins and Recent Evidence. *Trends in Cognitive Sciences*, *6*(1), 47–52. doi:10.1016/S1364-6613(00)01819-2

Baars, B. (2012). The Biological Cost of Consciousness. *Nature Proceedings*. doi:10.1038/npre.2012.6775

Baars, B., Franklin, S., Ramsoy, T. Z. (2013). Global Workspace Dynamics: Cortical "Binding and Propagation" Enables Conscious Contents. *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00200

Ballestar, E. (2010). Epigenetics Lessons from Twins: Prospects for Autoimmune Disease. Clinic. *Rev. Allergy. Immunol.*, *39*, 30–41. doi:10.1007/s12016-009-8168-4

Banadettini, P. A. (2009). What's New in Neuroimaging Methods? AnnNY. Acad. Sci., *1156*, 260–293. doi:10.1111/j.1749-6632.2009.04420.x

Barrett, A. B., Seth, A. K. (2011). Practical Measures of Integrated Information for Time-Series Data. *PLoS Computational Biology*, *7*(1). doi:10.1371/journal.pcbi.1001052

Bayne, T. (2009). Consciousness. In J. Symons, P. Calvo (Eds.), *The Routledge Companion to Philosophy of Psychology* (pp. 477–94). New York: Routledge.

Bayne, T., Hohwy, J., Owen, A. M. (2016). Are There Levels of Consciousness? *Trends in Cognitive Science*, *20*(6). doi:10.1016/j.tics.2016.03.009

Bisenius, S., Trapp, S., Neumann, J., Schroeter, M. L. (2015). Identifying Neural Correlates of Visual Consciousness With ALE Meta-Analyses. *Neuroimage*, *122*, 177–87. doi:10.1016/j.neuroimage.2015.07.070

Block, N. (1995). On Confusion About a Function of Consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–287.

Bossdorf, O., Richards, C. L., Pigliucci, M. (2008). Epigenetics for Ecologists. *Ecology Letters*, *11*, 106–115. doi:10.1111/j.1461-0248.2007.01130.x

Brook, A. (2008). Terminology in Consciousness Studies. Retrieved from: http://www.ym.edu.tw/assc12/tutorials.html#02

Carruthers, P. (1998). Animal Subjectivity. *Psyche*, *4*(3).

Carruthers, P. (2018). Comparative Psychology Without Consciousness. *Consciousness and Cognition*, *63*, 47–60. doi:10.1016/j.concog.2018.06.012

Carruthers, P., (2016). Higher-Order Theories of Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from: https://plato.stanford.edu/archives/fall2016/entries/consciousness-higher

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., …, Massimini, M. (2013). A Theoretically Based Index of Consciousness

Independent of Sensory Processing and Behavior. *Science Translational Medicine*, *5*(198), 198ra105.

Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., …, Massimini, M. (2016). Stratification of Unresponsive Patients by an Independently Validated Index of Brain Complexity. *Ann Neurol.*, *80*(5), 718–729.

Chalmers, D. (1995). Facing up to the Problem of Consciousness. *Journal of Consciousness Studies*, *3*, 200–219

Chalmers, D. (1996). The Conscious Mind: in Search of a Fundamental Theory. Oxford: Oxford University Press.

Cohen, M. A., Dennett, D. C. (2011). Consciousness Cannot Be Separated From Function. *Trends in Cognitive Science*, *15*, 358–364. doi:10.1016/j.tics.2011.06.008

Crane, T. (2000). The Origins of Qualia. In T. Crane, S. Patterson (Eds.), *The History of the Mind-Body Problem* (pp. 169–94). London: Routledge.

Crick, F., Koch, C. (2003). A Framework for Consciousness. *Nature Neuroscience*, *6*(2), 119–126.

Damasio, A. (1999). *The Feeling of What Happens: Body, Emotion and the Making of Consciousness*. London: Vintage.

Dehaene, S., Kerszberg, M., Changeux, J. P. (1998). A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks. *Proc. Natl. Acad. Sci. USA*, *95*, 14529–14534.

Dehaene, S., Naccache, L. (2001). Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework. *Cognition*, *79*, 1–37.

Dehaene, S., Changeux, J.-P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, *70*, 200–27. doi:10.1016/j.neuron.2011.03.018

Dehaene, S., Lau H., Kouider, S., (2017). What Is Consciousness, and Could Machines Have It? *Science*, *358*(6362), 486–492. doi:10.1126/science.aan8871

Dennett, D. C. (1988). Quining Qualia. In A. Marcel, E. Bisiach (Eds.), *Consciousness in Modern Science* (pp. 42–77). Oxford: Oxford University Press.

Earl, B. (2014). The Biological Function of Consciousness. *Frontiers in Psychology*, *5*(697). doi:10.3389/fpsyg.2014.00697

Edelman, G. (2003). Naturalizing Consciousness: A Theoretical Framework. *Proceedings of the National Academy of Sciences*, *100*(9), 5520–5524. doi:10.1073/pnas.0931349100

Edelman, G., Tononi, G. (2000). Re-Entry and the Dynamic Core: Neural Correlates of Conscious Experience. In T. Metzinger (Ed.), *Neural Correlates of Consciousness* (pp. 139–151). Cambridge, MA: MIT Press.

Edelman, D., Seth, A. (2009). Animal Consciousness: A Synthetic Approach. *Trends in Neuroscience*, *9*, 476–84. doi:10.1016/j.tins.2009.05.008

Edelman, G., Gally J. A., Baars, B. (2011). Biology of Consciousness. *Frontiers in Psychology*, *2*(4). doi:10.3389/fpsyg.2011.00004

Farisco, M., Evers, K. (2017). The Ethical Relevance of the Unconscious. *Philosophy, Ethics, and Humanities in Medicine*, *12*(11). doi:10.1186/s13010-017-0053-9

Feinberg, T. E., Mallatt, J. (2013). The Evolutionary and Genetic Origins of Consciousness in the Cambrian Period Over 500 Million Years Ago. *Frontiers in Psychology*, 4(667). doi:10.3389/fpsyg.2013.00667

Feinberg, T. E., Mallatt, J. (2016). The Nature of Primary Consciousness. A New Synthesis. *Consciousness and Cognition*, *43*, 113–127. doi:10.1016/j.concog.2016.05.009

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., Rees, G. (2010). Relating Introspective Accuracy to Individual Differences in Brain Structure. *Science*, 329(5998), 1541–1543. doi:10.1126/science.1191883

Fleming, S. M., Lau, H. C., (2014). How to Measure Metacognition. *Front. Hum. Neurosci.*, *8*(443). doi:10.3389/fnhum.2014.00443

Fraga, M. (2005). From the Cover: Epigenetic Differences Arise During the Lifetime of Monozy-Gotic Twins. *Proceedings of the National Academy of Sciences*, *102*(30), 10604–10609. doi:10.1073/pnas.0500398102

Freund, J., Brandmaier, A. M., Lewejohann, L., Kirste, I., Kritzler, M., Krüger, A., et al. (2013). Emergence of Individuality in Genetically Identical Mice. Science 340:6133, 756–759. doi:10.1126/science.1235294

Frith, C. D. (2011). What Brain Plasticity Reveals About the Nature of Consciousness: Commentary. *Frontiers in Psychology*, 2(87). doi:10.3389/fpsyg.2011.00087

Gennaro, R. (2005). The HOT Theory of Consciousness: Between a Rock and a Hard Place. *Journal of Consciousness Studies*, *12*, 3–21.

Giacino, J. T. (2005). The Minimally Conscious State: Defining the Borders of Consciousness. *Progress in Brain Research*, *150*, 381–95. doi:10.1016/S0079-6123(05)50027-X

Griffin, D. R. (2001). *Animal Minds: Beyond Cognition to Consciousness*. Chicago, IL: University of Chicago Press.

Heavey, C. L., Hurlburt, R. T. (2008). The Phenomena of Inner Experience. *Consciousness and Cognition*, *17*(3), 798–810. doi:10.1016/j.concog.2007.12.006

Hesselmann, G., Moors, P. (2015). Definitely Maybe: Can Unconscious Processes Perform the Same Functions as Conscious Processes? *Front. Psychol.*, *6*(584). doi:10.3389/fpsyg.2015.00584

Hassin, R. R. (2013). Yes It Can: On the Functional Abilities of the Human Unconscious. *Perspect. Psychol. Sci.*, *8*, 195–207. doi:10.1177/1745691612460684

Hohwy, J. (2009). The Neural Correlates of Consciousness. New Experimental Approaches Needed? *Consciousness and Cognition*, *18*, 428–38. doi:10.1016/j.concog.2009.02.006

Hollande, O. (Ed.). (2003). *Machine Consciousness*. Exeter: Imprint Academic.

Irvine, E. (2012) *Consciousness as a Scientific Concept: A Philosophy of Science Perspective*. Dordrecht: Springer.

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly*, *32*, 127–36.

Jonkisz, J. (2009). Świadomość i subiektywność – razem czy osobno. *Analiza i egzystencja*, *9*, 121–143.

Jonkisz, J. (2012). Consciousness: A Four-Fold Taxonomy. *Journal of Consciousness Studies*, *19*(11/12), 55–82.

Jonkisz, J. (2015). Consciousness: Individuated Information in Action. *Frontiers in Psychology*, *6*. doi:10.3389/fpsyg.2015.01035

Jonkisz, J. (2016). Subjectivity: A Case of Biological Individuation and an Adaptive Response to Informational Overflow. *Frontiers in Psychology*, *7*. doi:10.3389/fpsyg.2016.01206

Jonkisz, J., Wierzchoń, M., Binder, M. (2017). Four-Dimensional Graded Consciousness. *Frontiers in Psychology*, *8*. doi:10.3389/fpsyg.2017.00420

Kiefer, M. (2012). Executive Control Over Unconscious Cognition: Attentional Sensitization of Unconscious Information Processing. *Front. Hum. Neurosci. 6*. doi:10.3389/fnhum.2012.00061

Koch, C., Tononi, G. (2013). Can a Photodiode Be Conscious? *The New York Review of Books*. Retrieved from: http://www.nybooks.com/articles/archives/2013/mar/07/can-photodiode-be-conscious/

Koch, C., Massimini, M., Boly, M., Tononi, G., (2016). Neural Correlates of Consciousness: Progress and Problems. *Nat. Rev. Neurosci.*, 17(5), 307–21.

Kriegel, U. (2006). Consciousness: Phenomenal Consciousness, Access Consciousness, and Scientific Practice. In P. Thagard (Ed.), *Handbook of Philosophy of Psychology and Cognitive Science* (pp. 195–217). Amsterdam: North-Holland.

Kriegel, U. (2007). The Same-Order Monitoring Theory of Consciousness. *Synthesis Philosophica*, *2*, 361–384.

Lamme, V. A. (2006). Towards a True Neural Stance on Consciousness. *Trends in Cognitive Sciences*, *10/11*, 494–501. doi:10.1016/j.tics.2006.09.001

Lau, H., Rosenthal, D. (2011). Empirical Support for Higher-Order Theories of Conscious Awareness. *Trends. Cogn. Sci.*, *15*(8), 365–73. doi:10.1016/j.tics.2011.05.009

Laureys, S., Pellas, F., Van Eeckhout, P., Ghorbel, S., Schnakers, C., Perrin, F., …, Goldman, S. (2005). The Locked-in Syndrome: What Is It Like to Be Conscious but Paralyzed and Voiceless? *Progress in Brain Research*, *150*, 495–511. doi:10.1016/S0079-6123(05)50034-7

Legrand, D. (2007). Subjectivity and the Body: Introducing Basic Forms of Self-Consciousness. *Consciousness and Cognition*, *16*, 577–582. doi:10.1016/j.concog.2007.06.011

Levine, J. (1983). Materialism and Qualia: the Explanatory Gap. *Pacific Philosophical Quarterly*, *64*, 354–361.

Levine, J. (2001). Purple Haze: The Puzzle of Consciousness. Oxford and New York: Oxford University Press.

Lindahl, B. I. B. (1997). Consciousness and Biological Evolution. *Journal of Theoretical Biology*, *187*, 613–629. doi:10.1006/jtbi.1996.0394

Lycan, W. G. (1996). *Consciousness and Experience*. Cambridge, MA: MIT Press.

Marti, S., Kumar, K. H., Castellani, C. A., O'Reilly, R., Singh, S. M. (2011). Ontogenetic de Novo Copy Number Variations (CNVs) As a Source of Genetic

Individuality: Studies on Two Families With MZD Twins for Schizophrenia. *PloS One*, *6*(3). doi:10.1371/journal.pone.0017125

Merker, B. (2005). The Liabilities of Mobility: A Selection Pressure for the Transition to Consciousness in Animal Evolution. *Conscious. Cogn.*, *14*, 89–114.

Metzinger, T. (Ed.). (2000). *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. Cambridge, MA: The MIT Press/A Bradford Book.

Middlebrooks, P. G., Sommer, M. A. (2012). Neuronal Correlates of Metacognition in Primate Frontal Cortex. *Neuron*, *75*, 517–30. doi:10.1016/j.neuron.2012.05.028

Migicovsky, Z., Kovalchuk, I. (2011). Epigenetic Memory in Mammals. *Front. Gene.*, *2*(28). doi:10.3389/fgene.2011.00028

Morsella, M. (2005). The Function of Phenomenal States: Supramodular Interaction Theory. *Psychological Review*, *112*(4), 1000–1021. doi:10.1037/0033-295X.112.4.1000.PMDI16262477.

Morin, A. (2006). Levels of Consciousness and Self-Awareness. *Consciousness and Cognition*, *15*, 358–371.

Mudrik, L., Breska, A., Lamy D., Deouell, L. Y. (2011). Integration Without Awareness: Expanding the Limits of Unconscious Processing. *Psychological Science*, *22*(764). doi:10.1177/0956797611408736

Mudrik, L., Faivre, N., Koch, C. (2014). Information Integration Without Awareness. *Trends in Cognitive Science*, *18*(9), 488–496. doi:10.1016/j.tics.2014.04.009

Nagel, T. (1974). What Is It Like to Be a Bat? *Philosophical Review*, *83*, 435–451.

Neisser, J. (2006). Unconscious Subjectivity. *Psyche*, *12*(3). Retrieved from: http://www.theassc.org/files/assc/2642.pdf

Neisser, J. (2015) *The Science of Subjectivity*, London: Palgrave Macmillan. doi:10.1057/9781137466624

Northoff, G., Musholt, K. (2006). How Can Searle Avoid Property Dualism? Epistemic-Ontological Inference and Autoepistemic Limitation. *Philosophical Psychology*, *19*(5), 1–17.

Noë, A., Thompson, E. (2004). Are There Neural Correlates of Consciousness? *Journal of Consciousness Studies*, *11*(1), 3–28.

Overgaard, M. (Ed.). (2015). *Behavioural Methods in Consciousness Research*. Oxford: Oxford University Press.

Overgaard, M., Sandberg, K. (2012). Kinds of Access: Different Methods for Report Reveal Different Kinds of Metacognitive Access. *Philosophical Transactions of the Royal Society B*, *367*, 1287–1296. doi:10.1098/rstb.2011.0425

Overgaard, M., Rote, J., Mouridsen, K., Ramsoy, T. Z. (2006). Is Conscious Perception Gradual or Dichotomous? A Comparison of Report Methodologies During a Visual Task. *Consciousness and Cognition*, *15*, 700–708.

Overgaard, M., Timmermans, B., Sandberg, K., Cleeremans, A. (2010). Optimizing Subjective Measures of Consciousness. *Consciousness and Cognition*, *19*, 682–684. doi:10.1016/j.concog.2009.12.018

Palmer, T. D., Ramsey, A. K. (2012). The Function of Consciousness in Multisensory Integration. *Cognition*, *125*, 353–364. doi:10.1016/j.cognition.2012.08.003

Pally, R. (2005). Non-Conscious Prediction and a Role for Consciousness in Correcting Prediction Errors. *Cortex*, *41*, 643–62.

Pareira, A., Ricke, H. (2009). What Is Consciousness? Towards a Preliminary Definition. *Journal of Consciousness Studies*, *16*(5), 28–45.

Pfefferbaum, A., Sullivan, E.V., Carmelli, D. (2004). Morphological Changes in Aging Brain Structures Are Differentially Affected by Time-Linked Environmental Influences Despite Strong Genetic Stability. *Neurobiology of Aging*, *25*, 175–183. doi:10.1016/S0197-4580(03)00045-9

Pfeifer, R., Iida, F., Lungarella, M. (2014). Cognition From the Bottom Up: On Biological Inspiration, Body Morphology, and Soft Materials. *Trends in Cognitive Science*, *18*(8), 404–13. doi:10.1016/j.tics.2014.04.004

Pierre, J. (2003). Intentionality. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Retrieved from: http://plato.stanford.edu/entries/intentionality/#9

Rochat, P. (2011). The Self as Phenotype. *Consciousness and Cognition*, *20*(1), 109–19. doi:10.1016/j.concog.2010.09.012

Rosenthal, D. (1986). Two Concepts of Consciousness. *Philosophical Studies*, *49*, 329–359.

Sahraie, A., Hibbard P. B., Trevethan C. T., Ritchie K. L., Weiskrantz, L. (2010). Consciousness of the First Order in Blindsight, *PNAS*, *107*(49), 21217-21222. doi:10.1073/pnas.1015652107

Sanborn, A. N., Chater, N. (2016). Bayesian Brains Without Probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. doi: 10.1016/j.tics.2016.10.003

Schnakers, C. (2008). A French Validation Study of the Coma Recovery Scale-Revised (CRS-R). *Brain Injury*, *22*(10), 786–792. doi:10.1080/02699050802403557

Schnakers, C., Majerus, S., Goldman, S., Boly, M., Van Eeckhout, P., Gay, S., …, Laureys, S. (2008). Cognitive Function in the Locked-in Syndrome. *J. Neurol.*, 255(3), 323–30. doi:10.1007/s00415-008-0544-0

Searle, J. (1992). *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Searle, J. (2000). Consciousness. *Annual Review of Neuroscience*, *23*, 557–578.

Sergent, C., Dehaene, S. (2004). Is Consciousness a Gradual Phenomenon? Evidence for an All-or None Bifur-Cation During the Attentional Blink. *Psychological Science*, *15*(11), 720–729.

Seth, A. K., (2008). Causal Networks in Simulated Neural Systems. *Cognitive Neurodynamics*, *2*, 49–64.

Seth, A. K. (2009). Functions of Consciousness. In W. P. Banks (Ed.), *Encyclopedia of Consciousness* (pp. 279–293). Amsterdam: Elsevier/Academic Press.

Seth, A. K., Baars, B., and Edelman, D. (2005). Criteria for Consciousness in Humans and Other Mammals. *Consciousness and Cognition*, *14*(1), 119–139.

Seth, A. K., Friston, K. J., (2016). Active Interoceptive Inference and the Emotional Brain. *Phil. Trans. R. Soc. B*, *371*(1708), 20160007. doi:10.1098/rstb.2016.0007

Shimono, M., Mano, H., and Niki, K. (2012). The Brain Structural Hub of Interhemispheric Information Integration for Visual Motion Perception. *Cerebral Cortex*, *22*, 337–344. doi:10.1093/cercor/bhr108

Smith, E., Delargy, M. (2005). Locked-in Syndrome. *BMJ*, *330*, 406–9. doi:10.1136/bmj.330.7488.406

Swaddle, J. P., Cathey, M. G., Cornell, M., Hopkinton, B. P. (2005). Socially Transmitted Mate Preferences in a Monogamous Bird: A Non-Genetic Mechanism of Sexual Selection. *Proceedings. Biological sciences / The Royal Society*, *272*(1567). doi:10.1098/rspb.2005.3054

Teasdale, G., Jennett, B. (1974). Assessment of Coma and Impaired Consciousness. A Practical Scale. *Lancet II*, 81–86.

Theeuwes, J. (2010). Top-Down and Bottom-up Control of Visual Selection. *Acta Psychol (Amst)*, *135*(2), 77–99. doi:10.1016/j.actpsy.2010.02.006

Timmermans, B., Cleeremans, A. (2015). How Can We Measure Awareness? An Overview of Current Methods. In M. Overgaard (Ed.), *Behavioural Methods in Consciousness Research* (pp. 21–46). Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199688890.003.0003

Tononi, G. (2004). An Information Integration Theory of Consciousness. *BMC Neuroscience*, *5*(42). doi:10.1186/1471-2202-5-42

Tononi, G. (2008). Consciousness as Integrated Information: A Provisional Manifesto. *The Biological Bulletin*, *215*(3), 216–42. doi:10.2307/25470707

Tononi, G. (2010). Information Integration: Its Relevance to Brain Function and Consciousness. *Archives Italiennes de Biologie*, *148*, 299–322.

Tononi, G., Koch, C. (2014). Consciousness: Here, There but Not Everywhere. *Phil. Trans. R. Soc. B*, *370*(1668). doi:10.1098/rstb.2014.0167

Torrance, S., Clowes, R., Chrisley, R. (2007). Machine Consciousness Embodiment and Imagination. *Journal of Consciousness Studies*, *14*(7), 7–14.

Torrance, S. (2009) Contesting the Concept of Consciousness. *Journal of Consciousness Studies*, *16*(5), 111–126.

Tsuchiya, N., Wilke, M., Frässle, S., Lamme, V. A. F. (2015) No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Trends Cogn. Sci. 19*(12), 757–770. doi:10.1016/j.tics.2015.10.002

van Gaal, S., and Lamme, V. A. F. (2012). Unconscious High-Level Information Processing: Implication for Neurobiological Theories of Consciousness. *Neuroscientist*, *18*, 287–301. doi:10.1177/1073858411404079

van Gulick, R. (2018) Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from: https://plato.stanford.edu/archives/spr2018/entries/consciousness/

Velmans, M. (2009). How to Define and How Not to Define Consciousness. *Journal of Consciousness Studies*, *16*(5), 139–156.

Valizadeh, S. A., Liem, F., Mérillat, S., Hänggi, J., Jäncke, L. (2018). Identification of Individual Subjects on the Basis of Their Brain Anatomical Features. *Scientific Reports*, *8*(5611). doi:10.1038/s41598-018-23696-6.

Wierzchoń, M., Asanowicz, D., Paulewicz, B., Cleeremans, A. (2012). Subjective Measures of Consciousness in Artificial Grammar Learning Task. *Consciousness and Cognition*, 21(3), 1141–53. doi:10.1016/j.concog.2012.05.012

Wierzchoń, M., Paulewicz, B., Asanowicz, D., Timmerman, B., Cleeremans, A., (2014). Different Subjective Awareness Measures Demonstrate the Influence of Visual Identification on Perceptual Awareness Ratings. *Consciousness and Cognition*, *27*, 109–120. doi:10.1016/j.concog.2014.04.009

Windey, B., Gevers, W., Cleeremans, A. (2013). Subjective Visibility Depends on Level of Processing. *Cognition*, *129*(2), 404–9. doi:10.1016/j.cognition.2013.07.012

Windey, B., Vermeiren, A., Atas, A., Cleeremans, A. (2014). The Graded and Dichotomous Nature of Visual Awareness. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *369*(1641). doi:10.1098/rstb.2013.0282

Zhou, L., He, Z. J., Ooi, T. L. (2013). The Visual System's Intrinsic Bias and Knowledge of Size Mediate Perceived Size and Location in the Dark. *J. Exp. Psychol. Learn. Mem. Cogn.*, *39*(6). doi:10.1037/a0033088

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXV, nr 1 (2021), s. 35–54
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxv1.03
Article

BARBARA TOMCZYK [*]

KNOWER AT RISK: UPDATING EPISTEMOLOGY IN THE LIGHT OF ENHANCED REPRESENTATIONS

SUMMARY: The epistemological consequences of the increasing popularity of artificial cognitive enhancements are still confined to the margins of philosophical exploration, with priority given instead to ethical problems requiring urgent practical solutions. In this paper, I examine the less popular, yet still important, problem of the threats to which the very knowledge-forming process is exposed when its subject uses artificial cognitive enhancers. The theory of knowledge I call upon is borrowed from virtue epistemologists who, together with proponents of active externalism, seek to define the conditions that will protect artificially enhanced agents from a loss of epistemic agency. I invoke three such conditions (authenticity, integration and reciprocal causation), rejecting the last one. Incorporating active externalism into virtue epistemology points to the possibility of treating extended systems, composed of humans and artifacts, as extended subjects of knowledge. In the final part, however, I present two arguments against such an extension of epistemic agency.

KEYWORDS: cognitive enhancement, virtue epistemology, active externalism, extended cognitive system, epistemic agency.

Introduction

Knowledge is surely one of the most desirable of goods. It is considered a source of power and prosperity; its possession is rewarded and the lack of it rebuked. Modern technological developments have enabled the production of

[*] Maria Curie-Skłodowska University, Faculty of Philosophy and Sociology. E-mail: barbara.tomczyk@mail.umcs.pl. ORCID: 0000-0001-8145-7755.

artifacts supporting the acquisition of knowledge on a previously unprecedented scale—something that has inspired bold ideas about future possibilities. The most enticing of these is that of learning effortlessly and immediately, as seen in the 1999 movie *The Matrix*, where the acquisition of knowledge (of how to practice kung fu and operate a helicopter) by attaching electrodes to the agent's head seems to bypass all natural cognitive mechanisms. However, that which arouses fascination and envy among viewers of this movie raises concerns amongst epistemologists. Can beliefs or skills gained without effort embody the highest epistemic value? Can someone involved in this scenario be considered a genuine subject of knowledge? These doubts are of particular concern to virtue epistemologists, who define knowledge in terms of the cognitive achievement that the agent has attained using his or her own cognitive faculties, and for which he or she deserves credit. The need to reconcile the solutions of virtue epistemology with current and anticipated technological challenges has prompted precise analyses of the conditions that the agent's belief (or skill) must fulfill in order to be considered an instance of their knowing something. Although I will not be presenting any detailed analysis of these conditions, I will draw attention to them in order to answer the questions I myself find most pressing in the context of enhanced epistemic agency: Why, how, and from what should we protect the subject of knowledge that uses artificial cognitive enhancements? These questions are based on certain assumptions that I will be analyzing in subsequent parts of the article. To begin with, I indicate the particular understanding of epistemic agency that I intend to adopt. I borrow this from John Sosa, Ernest Greco and Duncan Pritchard—the founders of virtue epistemology. I also explicate the very need to protect agency against such threats. Next, I will substantiate the assumption about the existence of threats posed by the use of cognitive artifacts, and will point to the strategies for defense employed by virtue epistemologists. In the final part, I will consider the proposal of extending epistemic agency to include the entire extended cognitive system (consisting in each case of a human being and an artifact) that could potentially protect the agent's representations from the negative impact of using artifacts. Such a strategy seems to be an obvious consequence of including Clark and Chalmers' thesis of active externalism within the theory of knowledge under discussion. Nevertheless, I will show that this proposal does not achieve its intended goal, and therefore does not justify the introduction of the concept of an extended agent into epistemology.

Knowledge as a Cognitive Achievement and Its Subject

Knowledge is a mental representation of a special kind. Knowledge-that, which is of special interest to epistemologists, is a kind of belief: namely, an assertive attitude towards a given judgment. To count as knowledge, belief must meet additional requirements, of a kind that have been the focus of lively discussion amongst epistemologists ever since antiquity. For present purposes, I shall accept the conditions imposed on belief by proponents of virtue epistemology in the form of an externalist and reliabilistic theory of knowledge: one that introduces the concept of cognitive ability to reliabilism. Reliabilism itself states that the subject has a justified belief if, and only if, it is the product of a reliable cognitive process: i.e., one that in most cases leads to true belief (Goldman, 1979).[1] Virtue epistemologists have pointed out that this is an insufficient condition for knowledge, and have illustrated their point with many counterexamples.[2] Here, I present the most popular of them, which will also prove useful later on:

> TRUETEMP: Suppose a person, whom we shall name Mr. Truetemp, undergoes brain surgery by an experimental surgeon who invents a small device which is both a very accurate thermometer and a computational device capable of generating thoughts. The device, call it a tempucomp, is implanted in Truetemp's head so that the very tip of the device, no larger than the head of a pin, sits unnoticed on his scalp and acts as a sensor to transmit information about the temperature to the computational system in his brain. This device, in turn, sends a message to his brain causing him to think of the temperature recorded by the external sensor. Assume that the tempucomp is very reliable, and so his thoughts are correct temperature thoughts. All told, this is a reliable belief-forming process. Now imagine, finally, that he has no idea that the tempucomp has been inserted in his brain, is only slightly puzzled about why he thinks so obsessively about the temperature, but never checks a thermometer to determine whether these thoughts about the temperature are correct. He accepts them unreflectively, another effect of the tempucomp. Thus, he thinks and accepts that the temperature is 104 degrees. It is. Does he know that it is? (Lehrer, 1990, pp. 162–163)

The intuitive answer to the above question is "no". Thus, not every reliable belief-forming process leads to knowledge. According to virtue epistemologists, knowledge must derive from cognitive ability: i.e., the correctness of a known true belief must be due to the manifestation of a cognitive ability. Truetemp's belief derives from a reliable process, but not from any cognitive ability of his, and therefore he does not know (Greco, 2010; Pritchard, 2010). In order to assign Truetemp knowledge, the belief-forming process would have to have been appropriately integrated into his cognitive architecture, so that the belief would be the result of his cognitive abilities. Only after this condition is met can Truetemp, or any other agent, be considered a subject of epistemic credit and responsibility. The agent's cognitive character consists of all his cognitive abilities, both innate and acquired. What should be especially emphasized when discussing this theory of knowledge is the importance that its proponents attach to the properties of the belief-formation process itself. That is to say, this process cannot be truth-conducive through sheer luck, and cannot consist solely in the use of other people's cognitive abilities. Knowledge must be a product of the

[1] For detailed discussion of Goldman's theory of knowledge, see the present author's book-length study (Trybulec, 2012).

[2] Among the most influential virtue epistemologists, we should mention Ernest Sosa (1988; 2007), John Greco (1999), and Duncan Pritchard (2006).

cognitive abilities of its subject: only then does he or she own this special kind of representation—thus being responsible for it. What is important, moreover, is that the subject of knowledge does not have to be aware of the reliability of the processes resulting in this special kind of representation, or the extent to which they are integrated with his or her cognitive character. Virtue epistemologists are epistemic externalists, so they accept that the subject of knowledge need not know the way in which this epistemically valuable belief is formed. In the following, however, I intend to point out that this externalism has to be suspended when the agent, in order to solve a cognitive task, decides to go beyond his or her natural abilities and employ some artifact.[3]

There are two reasons why I refer to virtue epistemology when examining the influence of artificial cognitive enhancements on the process of acquiring knowledge. First, I recognize that its proponents have proposed an extremely insightful analysis of knowledge, presenting convincing solutions to many classic problems relating to this. Secondly, it is a theory that is constantly developing, whose proponents are actively engaged in upgrading previous solutions in the light of new cognitive phenomena and the philosophical concepts needed to explain them. Among such phenomena are artifacts that not only improve the natural cognitive processes, but also may, in the near future, enable the achievement of a cognitive goal that completely bypasses them. Yet the enthusiasm generated by such a vision is overshadowed by doubts as to whether such a process could be considered to represent a success on the part of the agent, such that he or she could be given credit for it. The growing popularity of artificial cognitive enhancements risks a blurring of epistemic responsibility and a decline in the value of knowledge—in which the latter may eventually cease to be a desirable achievement. Below, I will indicate in which cases of the use of artifacts the threat to cognitive achievement is the most real.

Cognitive Enhancements and Artificial Representations

The purpose of using cognitive enhancements is to quickly and effectively acquire knowledge, both propositional and procedural. Such enhancements include, in the broadest sense, any method that has the effect of improving the functioning of the human cognitive system. They can be divided into natural ones, such as learning, meditation and mnemonics, and artificial ones, which include the use of pharmacology, artificial intelligence and genetic modifications. In the narrow sense I am referring to in this paper, enhancement—as opposed to therapy such as is used to combat the effects of a neurological disease or injury—aims at improving the cognitive abilities of a healthy person. The improvement in question concerns both the receptivity of the human sensory apparatus and intellectual efficiency as this relates to memory, intelligence and creativity,

[3] For a more detailed discussion of virtue epistemology and its application to the study of extended cognitive systems, see the author's book (Trybulec, 2017).

and even to control over emotion, mood and desire (Sandberg, Bostrom, 1993). The use of artifacts that are external to the human body does not raise as many doubts as direct stimulation of the neural system. Thus, the ethical and epistemological discussion focuses mainly on cases of the second type, even though external enhancements also represent an important area of epistemological research.

Direct stimulation of the neuronal processes responsible for specific cognitive states usually takes the form of psychoactive substances or implants placed in appropriate areas of the brain. Such enhancements, due to their immediate effect, are much more effective than external artifacts but, on the other hand, they can lead to unforeseen, long-lasting and not always desirable side effects. As for psychoactive substances, many of these are obtained from plants commonly used to enhance attention, memory and creativity. Such effects are caused by, among other things, caffeine, theine, guaranine and nicotine, yet it is doubtful whether their use can be considered an instance of the enhancement of cognitive processes through artifacts. Meanwhile, there is no such doubt in the case of such chemicals as nootropic and precognitive drugs. These pharmaceuticals are mainly used therapeutically to slow down the cognitive damage caused by Alzheimer's and Parkinson's, and to prevent attention-deficit hyperactivity disorder (ADHD). However, they are also applied as cognitive enhancers in healthy people, because they improve the functioning of neurotransmitters and neurons, and ensure better blood circulation in the brain. A popular cognitive enhancer with a therapeutic purpose is, for example, Modafinil. Above all, this is used in the treatment of narcolepsy and sleep apnea, but it also has properties sought after by healthy people, as it accelerates the learning process by strengthening memory and engendering increased concentration (Gunia, 2015). Among the known psychoactive drugs that show a capacity for the enhancement of creativity, self-esteem and the desire for self-improvement, Prozak, an antidepressant, should also be mentioned. A much more dangerous group of enhancers are narcotic substances such as amphetamines and their derivatives, which stimulate and increase concentration, but are also highly addictive.

Alongside chemical substances, the largest group of artificial cognitive enhancers are IT artifacts created as a result of the development of artificial intelligence. As far as external artifacts are concerned, most of these function as memory stores, data-mining analysis and visualization programs aimed at supporting processes of reasoning, imagining and decision making (Kisielnicki, 2008). Devices connected to the human body, or implemented inside it, enter into more proximate and often reciprocal causal relations with brain processes, and a person usually does not have as much control over their operation as in the case of external artifacts. An example of the feedback that occurs directly between brain neural activity and such an artifact would be the brain-computer interface. It can be initiated using an electroencephalogram or, more invasively, by attaching electrodes to the cortex of the brain (Vallabhaneni, Wang, He, 2005). One case of such an interface is furnished by the project presented in 2019 by the company Neuralink, which, although designed to help people with neurological injuries, is ultimately intended to provide cognitive enhancement of unimaginable power by directly connecting the human brain with artificial intelligence. The connection consists in installing sensors in the brain in the form of thin threads that read neuronal activity and transmit the signal to an implant placed behind the ear. The implant, in turn, should decode this signal and send it to the computer running the appropriate program. As a result, it would be possible to send commands to artificial intelligence and receive information from the latter directly just via thought. The question that arises in the context of the discussion about agency is that of how much control a person would have over the representations directly produced in their mind by such an enhancement. It is the degree of this control that determines whether beliefs implemented in this artificial way can be considered knowledge understood as an achievement. Admittedly, not every representation that acquires the status of knowledge arises as a result of a human being's conscious decisions: that is not the case, for example, where sensory representations are concerned. All such mental states should, nevertheless, be produced by the cognitive abilities that belong to the person in question. Only then does he or she own these mental states and constitute their subject. In the case of the brain-computer interface described above, it seems that the representation can be created artificially, bypassing the natural cognitive process (or at least a significant part of it that is running in the perceptual apparatus). Yet is there really something wrong with that? In the next section, I will seek to justify a positive answer to that question by spelling out what I take to be the most serious threats to epistemic agency that are related to the use of artificial cognitive enhancers.

Enhanced, Yet Autonomous?

A necessary condition for assigning any kind of (moral, legal, epistemic) responsibility for some action undertaken, and thus for the possibility of its evaluation in terms of whatever value it realizes, is the intellectual autonomy of its subject. An agent is intellectually autonomous if he or she is able to make decisions according to his or her own will, and exercises control over the actions to which they lead. Obviously, the use of cognitive aids does not, as such, pose a threat to cognitive autonomy. Indeed, relying on other people's knowledge and obtaining information from reliable sources are essential for cognitive success. There is, however, a threshold beyond which this success ceases to be creditable to the agent: the agent must rely on others and other sources of knowledge "up to the point that doing so would be at the expense of her own capacity for self-direction. And this makes intellectual autonomy, essentially, a virtue of self-regulation in the acquisition and maintenance of our beliefs" (Carter, 2020a, p. 2940). The boundary of autonomous agency is not determined arbitrarily, but results from analyses of cases such as THRUTEMP, which have led virtue epistemologists to formulate the already-mentioned necessary condition for counting as knowledge. To recall, they require true belief to be the result of the agent's use of his or her own cognitive abilities. Adam Carter, one of the leading contemporary virtue epistemologists, analyzes this condition in detail in the context of developing technological cognitive enhancements and when considering their impact on the knowledge-forming process (Carter, 2020c; forthcoming). Ultimately, he formulates a definition of autonomous belief, proposing a condition that is supposed to protect epistemic agency against possible threats from the use of the latest—or even just anticipated—technology. Before presenting this proposal, I will specify exactly what it is intended to protect the subject of knowledge against.

The discussion concerning the risk of using cognitive enhancements has mainly unfolded in the field of ethics, and has raised many important issues that call urgently for both solutions and appropriate regulative responses.[4] The problem of knowledge as addressed by those dealing specifically with ethical issues is most strongly related to the issue of agent autonomy. I will devote some attention to it, as it is the ground from which epistemological doubts have arisen.

It seems that supporting natural cognitive abilities through artifacts can only be beneficial. An agent is able to perform a given task faster or better, and sometimes its execution is simply impossible without the use of the relevant artifact. Intuitively, when it comes to identifying the agent *qua* initiator of the enhanced cognitive activity, the situation seems clear: it is a human being. Yet deeper reflection reveals a basis for doubt. If cognitive success depends on the use of an artifact without which it would not have happened, is it still the agent's achievement? The person using the artifact still remains the agent, as he or she is the initiator of the activity, but the resulting success does not seem to be entirely creditable to him or her. The intuition underlying the problem of authenticity can be expressed in the following question: To whom do we ascribe the greater cognitive achievement—the person who solves mathematical problems aided by nothing but their own memory, or the one who uses a calculator for this purpose? Everyday life shows that innate talent and skills that have been developed are valued more highly than the use of a cognitive enhancement, even where the persons involved achieve the same goal at the same time. It might seem that what we appreciate is the effort that a person using his or her own cognitive ability has to make to solve a given task, but this is not always the case. A genius can multiply three-digit numbers effortlessly, yet this does not earn him any less credit. What seems decisive for the decision to attribute achieved success is the agent's use of his or her own cognitive abilities, whether innate or developed. This is a kind of capital that is difficult to trade, and therefore has a special value. Naturally, by paying for a prestigious education one can acquire highly valued cognitive abilities, but the process is long and tedious compared to the immediate effects of some psychoactive substances or intracerebral implants.

The question about the agent's autonomy in the context of cognitive enhancement is therefore as follows: In a scenario where an agent uses an artifact to perform a given task, to whom should the achievement, and thus the epistemic

[4] Ethical considerations pertaining to cognitive enhancements have been explored by, among others, Jan-Christoph Bublitz (2013) and Walter Veit (2018).

responsibility, be ascribed? Can a person who checks the result of performing addition on a calculator be credited with adding numbers? It seems that in this latter situation no one can be credited with any achievement: the activity of adding numbers together simply did not occur, and there is no subject to which the success of the calculation can be attributed. The only action in this scenario is that of a human being checking the result in a calculator without calculating it. The agent who calculates the sum is the person who carries out addition in his or her head, or on a piece of paper—although the latter activity counts for less, as if the mere fact of aiding oneself in one's task with anything reduces the level of success. Imagine, though, a situation in which, after using Modafinil, a person performs a calculation in his or her head that he or she would not have been able to do without this enhancement. Thus, the agent does not exploit some process executed by an artifact (calculator), but rather employs an artifact (Modafinil) in order to perform a cognitive process that, if he or she had been more gifted or better educated, would have been achievable naturally, using just his or her own cognitive resources. Does such an enhancement raise similar doubts as the use of a calculator? From an ethical point of view, the use of this drug may still be questionable, in that it results in the playing field ceasing to be a level one between enhanced and unenhanced individuals. Epistemically, however, as I will show below, the situation is clear: the subject of calculation is the human being, and Modafinil does not shift the responsibility away from him or her—nor does it diminish his or her cognitive success.

Using a calculator or the Internet to perform cognitive tasks, while raising some questions about who should take the epistemic credit, does not undermine human agency. The person is still the initiator of cognitive activity, and chooses the method of achieving the goal. The real challenge epistemologists have to face is when cognitive enhancement disrupts the agent's identity, rendering the mental states that cause the action inauthentic. Here I am not referring to numerical identity, but rather, so to speak, to "being the same person" as before the enhancement: to the maintenance of psychological continuity with oneself—i.e., with one's own character. Only when the condition is met of identifying with one's enhanced mental states, feeling in control of them and having them as one's own, can the agent take epistemic responsibility for the actions they cause. If the cognitive enhancement is strong enough to disturb the sense of identity with one's "former self", if a person loses their sense of decision-making and exercising control over the actions in question, then their agency will be put in question, as will be the possibility of praising or blaming them for any possible success or failure. Such a situation may happen when, by directly affecting brain structure, the enhancement modifies representations that guide the agent in their actions, or the general dispositions and talents that define their personality. Changes of personality while retaining agent identity are of course possible, but they must be introduced in an appropriate manner over the course of a process of education, so as to allow for gradual assimilation. Rapid pharmacological or IT modifications are not properly coupled to the natural human cognitive mechanism, making it difficult to identify the subject of the enhanced actions (Fischer, 2000). An additional complication is introduced by those enhancements—currently mainly pharmacological ones—that result in emotional states that are positively evaluated by the agent and mistakenly assessed by the latter as forming a part of his or her psychological character. The agent, guided in his or her action by such enhanced emotions, has a sense of agency, decision-making and preservation of identity, but the mental states responsible for determining action are not authentic, and this suffices to undermine his or her epistemic agency.

The problem of the agent's autonomy and the authenticity of their mental states in the context of cognitive enhancement has been analyzed in great detail from the perspective of ethics and the philosophy of law by Bublitz and Merkel (2009). These authors point out that the real threat to agency arises in situations where the natural cognitive process has been replaced by a completely different mechanism: for example, by an implant placed in the brain that takes over some of the natural cognitive functions. As for the pharmacological enhancers in current use, these do not constitute such replacements, as their operation consists in the optimization or modification of already existing structures and neural connections. Hence, the actions that result from these changes are still effects of the functioning of the mechanism owned by the agent in question, allowing the latter to retain full-blooded agency. On the other hand, such enhancements may well be regarded by those committed to the use of traditional, longer-lasting methods, such as involve an element of self-denial, as effortless shortcuts that cannot count as genuine cases of achievement. Nevertheless, these intuitions, motivated by a sense of unfairness, do not affect the epistemic status of enhanced representations, which, after meeting the appropriate conditions, can constitute full-fledged cognitive achievements. According to Bublitz and Merkel, the most important of these conditions is a conscious decision to utilize the enhancement made by an agent who knows the expected results of its application or, if unfamiliar with them, is aware of the risk being taken. In other words, a person, in order to be a responsible subject of his or her mental states and actions, cannot be manipulated in a way that is completely beyond his or her conscious control. When this happens, he or she ceases to be the subject of the actions performed, and the resulting belief cannot be regarded as their own cognitive achievement.

Even if the above condition is met, and the agent's identity is secure, those focused primarily on ethical issues remain concerned by the fact that, in the near future, cognitive enhancements may well remove certain obstacles in the absence of which it no longer makes sense to speak of something having been achieved (Kass, 2004). By depriving a human being of the need to make an effort, they will erase an important aspect of his or her life: one that relates to pride, praise, winning and admiration, but also to failure, shame and humiliation. When a goal comes effortlessly, it ceases to be an achievement and becomes an emotionless, trivial action that is hard to praise or criticize. The essence of this problem is accurately presented by Michael Sandel:

> [A]s the role of the enhancement increases, our admiration for the achievement fades. Or rather, our admiration for the achievement shifts from the player to his pharmacist… This suggests that our moral response to enhancement is a response to the diminished agency of the person whose achievement is enhanced. The more the athlete relies on drugs or genetic fixes, the less his performance represents his achievement. (Sandel, 2012, pp. 25–26)

When there is no possibility of losing, when one knows the "cheat code" for a given game, it loses its sense, as winning ceases to be satisfying in that it no longer delivers the same thrill. In most of the tasks that a person undertakes, effort is a necessary condition for considering its completion an achievement. Moreover, systematic artifactual support of a kind that frees the agent from the necessity of making any cognitive effort threatens him or her with an increasing level of dependence that may subsequently lead to complete cognitive impotence in situations where this enhancement is unavailable. This is what drivers who make constant use of car satellite navigation experience when their device fails or is fully discharged. Their employment of the enhancement causes their ability to orient themselves effectively in relation to their surroundings to decline drastically, resulting in a loss of epistemic agency. To counteract this threat, virtue epistemologists seek to precisely pinpoint those situations in which the use of cognitive enhancements contributes to a loss of control and agency, and how to avoid this.

Autonomous Belief, Reciprocal Causation, and Integration as Conditions for Epistemic Agency

Epistemologists, and those working in the area of ethics, agree that the most serious threat to epistemic agency is related to the possibility of manipulating the agent's cognitive processes and mental states in a way that is beyond his or her control. When this happens, the right to freedom of thought may be violated. More specifically, such a situation occurs when the agent is supplied, without his or her knowledge, with representations in a way that completely bypasses his or her natural cognitive process (*acquisition manipulation*), or when autonomous representations are, without his or her knowledge, eradicated from his or her mind (*eradication manipulation*) (Carter, 2020b). As long as the mind was reduced to a Cartesian thinking substance, and the content of mental states was available only to the subject, the threat of thought manipulation amounted to mere theoretical speculation. Yet technological developments that may, in the near future, lead to an avalanche of artificial cognitive enhancements, have made it a practical possibility that urgently needs to be counteracted. Additionally, the mind has been "weakened" in its defense against manipulation by the increasingly influential idea that it may extend beyond the skull, and even beyond the agent's organism, in a way that involves processes occurring, and information states obtaining, in artifacts themselves. This idea, proposed by Andy Clark and David Chalmers under the label of "active externalism", indicates that in some cases of cognitive activity, a person is coupled with an external artifact in such a strong causal relationship (*continuous reciprocal causation*) that they co-constitute a single cognitive system (Clark, Chalmers, 1998). The physical realization base of cognitive processes, dispositional beliefs, or perceptual states may therefore extend beyond the safe, Cartesian "theater of the mind" into a publicly accessible world. Hence, given that some thoughts can be realized outside of the brain, they also need to be protected from the two types of manipulation mentioned above.

Adam Carter has carried out a highly detailed and insightful analysis of the condition that must be satisfied where autonomous belief is concerned, in order to serve as a protection in respect of artificially enhanced representations purporting to constitute knowledge. Here, I will only seek to the general contours of its overall outcome. In short, a belief will count as autonomous if, and only if, it has a compulsion-free history. This, in turn, will be the case if and only if the agent has not acquired the belief in a way that so bypasses or preempts his or her cognitive competences as to leave the agent improperly incapable of dispensing with that belief (Carter, 2020c). The subject of knowledge should, in other words, acquire a true belief as a result of using their own, unmanipulated cognitive abilities. If, however, these abilities are enhanced by some artifact, it should be properly integrated with the agent's cognitive character. This integration will be of a different nature to that which occurs in the process of acquiring new cognitive abilities or improving existing ones by methods of natural development. In the latter case, new dispositions do not have to be consciously accepted by the agent as is required in the scenario of an artifact being utilized. Virtue epistemologists, in collaboration with proponents of active externalism, have sought to explain how artificial enhancement can be integrated into the agent's cognitive character so that its use does not undermine their epistemic credit, and thus their knowledge.[5] In particular, they indicate two conditions that must be met for this to happen. First, according to the guidelines of Clark and Chalmers, the processes taking place within the agent and inside the artifact must be continuously linked via feedback loops. When this happens, the human being and the artifact form one system, in which the boundaries between organic and external processes are blurred, so that their separate study becomes futile. This kind of feedback only occurs if the enhancement is constantly present in the agent's life, easily and directly accessible, and applied uncritically, in a manner analogous to biological cognitive processes (Clark, Chalmers, 1998). Second, at some point in their life, the agent must have consciously incorporated the external enhancement into their cognitive abilities by accepting it as reliable (Pritchard, 2010).[6]

[5] Notably, the "Extended Knowledge Project", led by Duncan Pritchard, was undertaken at the University of Edinburgh from 2013 to 2015. As a part of this, virtue epistemologists (Pritchard, Carter) collaborated with proponents of active externalism (Clark, Palermos). The results obtained were published in book form (Carter, Clark, Kallestrup, Palermos, Pritchard, 2018).

[6] This condition is also present in Clark and Chalmers (1998), and in Rowlands (2010).

This requirement is illustrated by the Truetemp case described in the first paragraph of the present article. To remind readers, Truetemp, although he can determine the temperature in the room, has no knowledge of it, because this belief did not arise as a result of his cognitive abilities, but rather due to the operation of a device inserted into his brain. In order to attribute knowledge to Truetemp, it must be at least assumed that he knows the source of his true beliefs and has accepted them as reliable. Now let us consider another case. We may imagine that a scientist has installed a sensory substitution system in the body of a blind person without his or her knowledge. It is a device that converts information specific to the damaged sense modality into stimuli received by a working one. Would we consider the cognitive success caused by the operation of such a system to be the result of this person's use of his or her extended cognitive abilities? It seems not. Such a person is in the same epistemic scenario as Truetemp, because he or she has never consciously included a new competence into the framework of his or her cognitive system. They do not know the source of their reliability, and so could not be credited for the success in question.

The doubts that pertain to the influence of cognitive enhancements on epistemic agency do not therefore concern the sheer fact of their application, but rather their proper integration with the agent's cognitive system. At this point, it is worth emphasizing once again the difference between biological (natural) and extended (enhanced) cognitive processes. In the case of the former, the condition of consciously endorsing them as reliable and making a decision to use them does not have to be met in order for them to count as constitutive of the agent's cognitive character. This condition concerns only artificial cognitive enhancements used to improve biological processes. However, we should keep in mind that virtue epistemology typically adopts a reliabilistic stance towards knowledge. To remind readers, the epistemic status of a belief is determined, according to its proponents, by properties of the belief-forming process. Moreover, the agent need not be aware of these properties, and need not know whether the process is reliable or whether it meets other conditions proposed by virtue epistemologists. In this respect it is tantamount to an externalist theory of knowledge. Yet the above considerations pertaining to the need for conscious integration of cognitive enhancements with biological processes on the part of agents are internalistic in nature. In order to incorporate the process of manipulating an artifact into the framework of their cognitive systems, agents must, at some point in their lives, consciously acknowledge this enhancement as reliable, and embrace its continuous and unreflective utilization. In other words, an agent must know, or have known at some point in their life, the reasons underpinning the belief they now have as a result of using the relevant artifact. Hence, while working out the conditions governing knowledge for artificially enhanced agents, the virtue epistemologist must part company with the externalists and take up instead the position of some kind of internalist-reliabilistic hybrid.[7]

[7] I also develop this line of reflection in my book-length study (Trybulec, 2017).

Upgrading Epistemology With Active Externalism: Some Problems

All the conditions for the safe use of cognitive enhancement indicated in the previous paragraph focused on its integration with the agent's cognitive character. The most important challenge for epistemologists is to explain what, exactly, this integration is supposed to amount to. One answer, as I have already shown, is suggested by proponents of active externalism. Let me recall that, according to Clark and Chalmers, proper integration should consist of a continuous and reciprocal causal link between the agent's natural cognitive abilities and the processes taking place in the artifact itself. This means that the functioning of the former changes the operation of the latter, which in turn affects the former, and so on.[8] It is worth pausing for a moment here to reflect carefully on this. It will not take long before one realizes that the condition of reciprocal causal coupling appears too strong and difficult to fulfill when using some artificial enhancements. Is it possible, for example, to constitute such a dynamic system out of the conjunction of a human being with a psychoactive substance such as Modafinil? It seems that in the scenario of taking a pill, the causal relationship is one-sided and consists solely in the effect of the substance on the human nervous system, without feedback. The human being can, at most, monitor the changes taking place in his or her cognitive functioning and control the dose of the substance, but is not able, consciously or not, to change its impact on his or her cognitive character. Nevertheless, the condition of exerting control and retaining a sense of agency in the face of such an enhanced cognitive character is fulfilled, and it would be implausible to claim of such an artifact that it had taken epistemic responsibility away from the agent. The belief generated with the support of Modafinil is autonomous, and the agent has deliberately decided to incorporate this substance into the cognitive abilities responsible for this mental state. It seems, therefore, that the condition of reciprocal causal coupling, considered necessary by Clark and Chalmers for the existence of an extended system, is too strong when it comes to determining what counts as an epistemically safe utilization of a pharmacological artifact. In short, not every coupling between a human and an artifact that results in knowledge constitutes an extended cognitive system.

Active externalism seems to fall short of the hopes invested in it by epistemologists: the condition that it specifies, of an enhancement's having to be integrated with the agent's cognitive character, is not necessary for knowledge to be obtained through manipulation of the artifact. There is, moreover, a tension between active externalism and virtue epistemology, due to the internalist condition that requires the agent to consciously embrace the extended cognitive process as being reliable. That is to say, it does not favor the functionalist attitude that marks out supporters of active externalism in their dispute with those seeking to assert the importance of biologically determined prejudices ("bio-prejudices").

[8] The idea of mutual feedback as a necessary condition for epistemic subjectivity being enhanced by an artifact is analyzed in detail by Palermos (2014).

According to functionalists, the nature of the cognitive process (be it biological or artificial) is irrelevant to its knowledge-conducive function. Yet the intuitions extracted by virtue epistemologists by means of many thought experiments indicate the weakness of this position (Carter, 2013). Biological and artificially enhanced processes are not epistemically equivalent. As has already been noted, in order to incorporate the manipulation of artificial cognitive enhancement into the agent's cognitive character, the agent must consciously and freely decide about it, which he or she need not do in the case of such biological processes as we see manifested in our ordinary perceptual or rational faculties.

To maintain epistemic agency, the agent supporting himself or herself with some artifact should be concerned about its proper integration with their cognitive character. When deciding to use an artificial cognitive enhancement, they ought to be vigilant and attentive. The more thoroughly agents have familiarized themselves with how an artifact works, and how it affects their natural cognitive processes, the better protected they will be against manipulation or loss of control over the corresponding artificially enhanced process of belief-formation. Active externalism defines the conditions for an extended cognitive system whose cognitive processes are distributed and impossible to divide into the biological and the artificial. Yet, as was shown, not every use of an artifact in the knowledge-forming process constitutes such a system. Any such use, however, requires a person to consciously and freely accept the coupling between artificial enhancement and his or her natural cognitive processes, regardless of whether it be one-way or reciprocal. Hence, even in the case of very radical cognitive enhancement of the sort that is, at present, only part of a boldly anticipated future, maintaining the agent's cognitive autonomy is possible. Human cognitive dependence on technology is inevitable, but so long as epistemic vigilance is maintained it need not be detrimental to our epistemic agency. To reiterate, the possibility of assigning a cognitive achievement of sorts to the agent will be determined not so much by the type of enhancement utilized by the latter, but rather by the kind of influence it exerts on the agent and the degree of control the agent exercises over it.

Beyond Control

Even when all conditions for knowledge acquired with the support of artificial enhancement are met, epistemologists still have reservations. By way of concluding these considerations, I will point to the two areas of doubt that I consider the most serious. The internalist condition requiring the agent to consciously accept the impact of artificial enhancement on natural cognitive abilities significantly limits the technological possibilities for generating knowledge. That is, one cannot produce it by secretly installing a belief-forming implant, or administering a psychoactive substance to the agent. Of course, it is possible—albeit only theoretically, for the time being—to artificially and discreetly create in the agent's mind a representation with some appropriate content, but this will not count as knowledge from an epistemological point of view. Hence, the subject of knowledge seems to be protected from cognitive manipulation, yet the question arises as to how realistic and effective this protection is in practice. The consequences of using an enhancement, such as a psychoactive substance, can be somewhat unpredictable not only for the agent, but also even for specialists charged with controlling its use. Even if we assume that the agent is familiar with the nature of the influence exerted by a given substance, he or she may not be able to distinguish between his or her natural mental states and those produced by the enhancement itself. As a consequence, the agent loses control over the artifact and becomes susceptible to manipulation by other people, which leads to a loss of ownership of the resulting mental state. On the other hand, as was already indicated, after consciously incorporating enhancement into his or her cognitive character, the agent no longer needs to constantly control it. The artifact can become a part of the agent's cognitive system that works beyond the bounds of his or her consciousness. Were it not for the problematic internalistic condition that speaks in favor of "bio-prejudices", this would be an ideal scenario for adherents of active externalism. The extended cognitive system would function as a natural one and would not require any special treatment. The bad news, however, is that special treatment is indeed necessary—a point emphasized not only by virtue epistemologists, but also by the proponents of active externalism themselves. Clark and Chalmers give expression to this necessity by formulating four conditions for having a mental state (a belief) partly realized by an artifact (Clark, Chalmers, 1998), thus lending support—surely against their own intentions—to the thesis propounding the cognitive advantageousness of biological processes.

Another weak point when it comes to defending epistemic agency against the negative influence of cognitive enhancement concerns the authenticity of the mental states responsible for its control and the sense of ownership of the cognitive character that results. Carter's account of what is required in order to preserve the authenticity of belief draws attention to the necessity of using only the agent's cognitive abilities in the knowledge-forming process. Imagine, however, that the enhancement (be it a pharmaceutical one or an implant), though applied by the agent voluntarily, shapes his or her mental states responsible for the sense of control and agency. Assume, moreover, that the agent has agreed to such an influence, and—even more—that he or she has agreed to the artifact changing his or her identity (desires, emotions and beliefs). Are his or her mental states still authentic? It seems not, since they did not result from the agent's cognitive abilities. On the other hand, the agent has consciously and voluntarily incorporated the artifact into his or her cognitive character, making its processes his or her own. Actually, if there is reciprocal causation between natural and artificial processes, it is difficult to distinguish one from the other because they shape each other. Hence, it becomes impossible to determine whether a given mental state was triggered by the agent's cognitive abilities or by artificial processes that bypass his or her cognitive character. Even if this scenario represents no more than an audacious imagining of future possibilities, epistemologists surely need to prepare for it and be aware of any doubts about, or threats to, epistemic agency that it may bring on, even if they do not have ready solutions yet. Maybe, in the scenario just described, it would make sense to accept the proposal that, together with a human being, the artifact constitutes the agency of an extended system, or even that it comes to partly make up the subject of knowledge produced within such a system.[9] Having said this, while tempting, I myself do not consider this solution satisfactory.

Extending the realization base of epistemic agency to an artifact, while it may seem theoretically possible, does not, in my opinion, compel us to accept the thesis of an extended subject of knowledge. I would like to point out two reasons for such a verdict. Firstly, the subject of cognition bears epistemic responsibility for success or failure. Yet only a reflective system can be thus responsible. Such a system is distinguished by the ability to assess one's own mental states in terms of rationality and compliance with some adopted hierarchy of values. It also has the ability to make a free choice based on consideration of its possible consequences. In order to do that, an agent must have access to the contents of his or her mental states, be aware that they belong to himself or herself, realize that they derive from his or her own cognitive abilities, and be of the conviction that he or she controls them. Even if these conditions are met when aided by some cognitive enhancement, epistemic responsibility, which is associated with the apportioning of credit and blame, rests with the human and not with the human-plus-artifact. Surely, though, this is not the case when the sense of agency is created without one's knowledge or will, as it is when one's natural cognitive abilities are completely bypassed or manipulated, so that the condition of autonomy and cognitive integration is not met. In any other (non-pathological) case, the subject of knowledge will be the human being, because only he or she can be the object of epistemic evaluation—and of any reward or punishment associated with this.

One may wonder, nevertheless, whether it is at all possible for the realization base of epistemic agency to be extended without the agent itself also being so. Since mental states determining agency would be co-realized by artificial processes linked via reciprocal causation with natural ones, why not consider them states of the extended agent taken as a whole, and not just of one of its parts (i.e., the human being)? The first reason for not doing so has been outlined above, and concerns our intuitions and practices relating to the attribution of agency. At the same time, a theoretical grounding for this has been provided by Lynne R. Baker (2009), and this may itself be regarded as furnishing our second reason for confining epistemic agency to the human being within an extended cognitive system. While Baker's proposal concerns our understanding of the self in extended cognitive systems, it is not too much of a distortion to apply it to mental states in general. She refers to the division of reality into levels introduced by proponents

[9] The thesis of extended agency has been developed by, among others, Malafouris (2008). For its analysis and evaluation by the present author (Trybulec, 2020).

of nonreductive physicalism. Mental states, according to this stance, belong to the properties of a higher level of the cognitive system, and arise from a lower level, that of physical properties. The two types of properties have different characteristics. Physical properties, as opposed to mental ones, manifest themselves in space—within the agent's organism, or outside it. Higher-level systemic properties, such as agency, do not occupy space, so it is impossible to determine whether they are inside or outside the agent's body. As Baker argues, the fact that the social, linguistic, and physical environment plays a vital role in shaping the agent's mental states, and even partly realizes some of them, does not mean that the agent himself or herself is extended in any way. In other words, the agent's subpersonal states may consist in part of extra-biological elements that, by entering into complex causal relationships with one another, produce higher-level systemic properties such as beliefs, desires, and other mental states. The physical realization base of agency is in this case extended, but the agent itself is not, as the term "extended" applies only to physical properties. This observation seems to undermine the very thesis of the extended mind as put forward by Clark and Chalmers. However, I will not address that problem here. At this point, I would like instead to just focus on drawing the conclusion that artificial cognitive enhancements cannot take over some of the epistemic credit and responsibility from human beings, and therefore cannot share epistemic agency with them.

Concluding Remarks

The goal I set myself in this paper was to consider the epistemological consequences of the increasing popularity of artificial cognitive enhancements. Technological developments that are such as to allow for reasonable predictions as to their future mode of operation are of legitimate concern to philosophers studying the conditions of agency. The alarm has been raised primarily by those dealing with ethics, as the consequences of the increasing influence of artifacts on the human mind are linked to practical issues of social justice, and so demand urgent regulation. In the present article, though, I have sought to address another dimension of this phenomenon—the epistemological worries, which are less popular and therefore less frequently raised. I have pointed to scenarios in which the use of an artifact may deprive the agent of cognitive achievement, making him or her lose epistemic agency. I have also looked at three conditions for enhanced belief and knowledge (authenticity, reciprocal causation, and integration) that are suggested in the literature on this issue, and have dismissed one of them (reciprocal causation) as unnecessary. Despite my rejection of this condition, I find the collaboration of virtue epistemologists with supporters of active externalism to be most fruitful. The latter have certainly enriched epistemological considerations with their explanation of the relationship that unites a human being and an artifact into a single knowledge-forming (epistemic) system, and this suggests the possibility of treating such an object as an extended agent, where mental states such as knowledge, intentions and desires belong not just to the human being, but rather to the entire system. Such an approach would make it possible to solve the problem of what it means for human identity and agency to be distorted by an artifact that has nevertheless been correctly incorporated into the framework of the agent's cognitive character: in such cases, an artifact would co-constitute an instance of extended identity and agency—i.e., it would share these with the human being involved. On the other hand, in the final part of this paper, I have presented two arguments against such an extension of epistemic agency. Of these, the former refers to the close connection of agency with responsibility, while the latter invokes the concept of systemic properties that have different characteristics from their physical realization base.

As a consequence of the considerations pursued here, a doubt may arise as to why we should care about protecting epistemic agency at all. Is the dissolution of the subject of knowledge really something we should fight against? Well, yes! The decline of the epistemic agent entails a fading away of epistemic responsibility. That is to say, if there is no one to attribute a given achievement to, then no one can be responsible for either the cognitive success in question or its absence. Epistemologists are resolutely engaged in searching out the conditions for knowledge that will serve as its touchstone in every—even the most fantastic—scenario. These efforts, though, are not driven solely by theoretical ambitions. Doubts about the subject of knowledge resulting from enhanced cognitive abilities may already, in the near future, cause practical problems relating to the need to determine who should be praised or blamed for a given result. In this paper, I have also pointed to the problem of the reduction of cognitive effort, which becomes ever more serious, the more frequently and systematically people use enhancements. Both the lack of a need to demonstrate one's own skills and the lack of any risk of failure contribute to lowered self-esteem, as well as to a diminution in the sense of satisfaction associated with success and of anger connected with failure—both emotions that motivate self-improvement and development. All these doubts and concerns are sufficient reasons to care about the authenticity of our mental representations, and for taking seriously appeals for epistemic control and vigilance in the face of the rapid technological developments surrounding cognitive enhancements.

REFERENCES

Baker, L. (2009). Persons and the Extended Mind Thesis. *Zygon, 44*(3), 642–658.

Bublitz, J.-C., Merkel, R. (2009). Autonomy and Authenticity of Enhanced Personality Traits. *Bioethics, 23*(6), 360–374.

Bublitz, J.-C. (2013). My Mind is Mine!? Cognitive Liberty as a Legal Concept. In: E. Hildt, A. Francke (Eds.), *Cognitive Enhancement* (pp. 233–264). Dordrecht, New York: Springer.

Carter, A. (2013). Extended Cognition and Epistemic Luck. *Synthese, 190*(19), 4201–4214.

Carter, A., Clark, A., Kallestrup, J., Palermos, S. O., Pritchard, D. (Eds.). (2018). *Extended Epistemology*. Oxford: OUP.

Carter, A. (2020a). Intellectual Autonomy, Epistemic Dependence and Cognitive Enhancement. *Synthese, 197*, 2937–2961.

Carter, A. (2020b). Varieties of (Extended) Thought Manipulation. In: M. Blitz, C. Bublitz (Eds.), *The Future of Freedom of Thought: Liberty, Technology, and Neuroscience*. London: Palgrave Macmillan. Manuscript submitted for publication.

Carter, A. (2020c). Epistemic Autonomy and Externalism. In: K. Lougheed, J. Matheson (Eds.), *Epistemic Autonomy.* London: Routledge.

Clark, A., Chalmers, D. (1998). The Extended Mind. *Analysis, 58*(1), 7–19.

Fischer, J. (2000). Responsibility, History and Manipulation. *Journal of Ethics, 4*, 385–391.

Goldman, A. (1979). What is Justified Belief? In: G. S. Pappas, (Ed.), *Justification and Knowledge* (pp. 1–25). Dordrecht: Reidel.

Greco, J. (1999). Agent Reliabilism, *Philosophical Perspectives*, *13*, 273–296.

Greco, J. (2010). *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity*. Cambridge: CUP.

Gunia, A. T. (2015). Koncepcje wzmocnienia poznawczego. Próba definicji oraz przegląd metod. *Avant, VI*(2), 35–56.

Kass, L. R. (2004). *Life, Liberty and the Defense of Dignity: The Challenge for Bioethics*. San Francisco: Encounter Books.

Kisielnicki, J. (2008). *MIS – Systemy Informatyczne Zarządzania*. Warsaw: Placet.

Lehrer, K. (1990). *Theory of Knowledge*. London: Routledge.

Malafouris, L. (2008). At the Potter's Wheel: An Argument for Material Agency. In: C. Knappet, L. Malafouris (Eds.), *Material Agency. Towards a Non-Anthropocentric Approach* (pp. 19–36). New York: Springer.

Palermos, O. S. (2014). Knowledge and Cognitive Integration. *Synthese, 191*, 1931–1951.

Pritchard, D. (2006). *What is This Thing Called Knowledge?* New York: Routledge.

Pritchard, D. (2010). Cognitive Ability and the Extended Cognition Thesis. *Synthese, 175*, 133–151.

Rowlands, M. (2010). *The New Science of Mind*. Cambridge MA: The MIT Press (A Bradford Book).

Sandberg, A., Bostrom, N. (2006). Converging Cognitive Enhancements. In: W. S. Bainbridge, M. C. Roco (Eds.), *Annals of the New York Academy of Sciences* (1093, pp. 201–227). Oxford: Blackwell.

Sandel, M. J. (2012). *The Case against Perfection: What's Wrong with Designer Children, Bionic Athletes, and Genetic Engineering?* In: S. Holland (Ed.), *Arguing About Bioethics* (pp. 25–26). London: Routledge.

Sosa, E. (1988). Beyond Skepticism, to the Best of our Knowledge. *Mind, 97*, 153–89.

Sosa, E. (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowledge*. Oxford: Clarendon Press.

Trybulec, B. (2012). *Epistemologia znaturalizowana a normatywność*. Lublin: Wydawnictwo UMCS.

Trybulec, B. (2017). *Wiedza i jej podmiot w szerokich systemach poznawczych*. Warsaw: IFiS PAN.

Trybulec, B. (2020). Podmiot czy agent? Rozumienie podmiotowości w erze artefaktów poznawczych. *Filozofia i Nauka. Studia filozoficzne i interdyscyplinarne, 8*(2), 89–113.

Vallabhaneni, A., Wang, T., He, B. (2005). Brain-Computer Interface. In: B. He (Ed.), *Neural Engineering* (pp. 85–121). New York: Springer US.

Veit, W. (2018). Cognitive Enhancement and the Threat of Inequality. *Journal of Cognitive Enhancement, 2*, 404–410.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXV, nr 1 (2021), s. 55–72
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxv1.04

Article

MARIAM ARARATOVNA SARGSYAN[*]

METAPHOR IN SEMIOTICS: FOUNDATIONS, EMBODIMENTS, ANALYSIS

SUMMARY: A metaphor within the framework of semiotics can be embodied in various semiotic systems, which is a prerequisite for a multilateral, in-depth analysis of its generation and interpretation. The purpose of the article is the conceptualisation of metaphor in the framework of semiotics and analysis using methods of analogy and transference. One of the main problems of metaphor theory is to provide means to represent the process of metaphor generation for understanding the nature of the phenomenon. The use of the offered methods in metaphor generation and interpretation opens up a multifaceted understanding of the object under study.

KEYWORDS: metaphor, analogy, metaphorical transfer, metonymic series, coding.

1. Some Basic Characteristics of the Analysis of Metaphor as a Semiotic Sign

There are various approaches to the definition and analysis of metaphors. Most studies in the field of metaphors have focused on analysis in literary texts. This paper proposes to use the semiotic approach, with the help of which a comprehensive and more detailed analysis of the phenomenon under study is possible. Semiotics is the science of signs and sign systems involved in the communication process, which allows the analysis of a metaphor from the side of the one who generates it and the one who "consumes" it. One of the first people to define

[*] Southern Federal University, Faculty of Philosophy. E-mail: m.sargsyan93@gmail.com. ORCID: 0000-0002-0661-3700.

metaphor as a semiotic sign was Charles Sanders Peirce. According to the second trichotomy of signs proposed by Peirce, there are three types of signs concerning the object: an Icon, an Index, and a Symbol. An iconic sign represents an object mainly through similarity. A sign should be called hypoiconic if some additional substantive or interpreter is needed to represent an object. Hypoicons may be divided into three types: images which represent the relations, mainly dyadic; diagrams which are related to didactic relationships between parts of one object through similar relationships between their parts; and metaphors which represent the representative character of the sign by representing parallelism in something else (Peirce, 2000, pp. 200–202). According to the definition provided by Peirce, a metaphor is a hypoiconic sign rather than an iconic one because it is not based on the actual (literal) similarity of the significant and the signified, which implies the presence of certain interpreters for its understanding. The classification of hypoiconic signs indicates that the metaphor is not an image since it does not represent a direct (denotative) description of the primary qualities of an object. This is certainly true in the case of the metaphor "visual noise" (Rosengren, 2019, p. 88). The primary simple qualities of noise associated with hearing are defined in the metaphor through visual organs unusual for the perception of the object, and vice versa, vision is comprehended through the noise that does not directly represent it. Such comprehension of various things within one metaphorical formation is possible, since the metaphor is related to the universe of discourse, and by the provision of discursive registers and code parameters, it can be interpreted and understood (Sørensen, 2011, pp. 151–152). The metaphorical relationship between the various terms can be understood, since discourse and its inherent discursive registers allow one term to be embodied in another. The analysis is based on codes that establish some correspondence between the significant and the signified. The metaphor is not the second type of hypoiconic sign—a diagram—because it includes parts of various things based on the specific parallelism that it creates between them. The parallelism that is created by the metaphor can be described as the possibility of attributing some significant to a secondary signified, associated with the primary signified by similarity (Morris, 2001, pp. 121–122). This can be seen in the metaphor "visual noise" because the eyes register some photons of light reflected from objects, but we do not understand what they represent, it turns out that we look, but do not see, as in the case of noise, when we hear a set of chaotic sounds from which it is difficult to isolate something for perception. The metaphor, proceeding from the classification of Pierce, is a hypoiconic sign, which is similar to its object in some aspects based on the specific parallelism that exists between the signified and the significant.

A significant contribution to the study of the metaphorical sign was made by the philosopher Umberto Eco, who was engaged in the study of the functioning of metaphor and tools for its creation. He paid attention to the concept of metaphorical similarity as one of the possible grounds for creating a metaphor. Similarity, according to Eco, is characterised as replacing one term with another based on the relationship of semantic-positional similarity within the semantic system (Eco, 2005, pp. 137–138). Examples of such similarities can be found in the field of advertising as a semiotic system, in which metaphors of different purposes are often used. A group of Taiwanese scientists involved in the issue of visual metaphors in advertising proposed a classification of metaphors based on whether or not the product's likeness is incorporated into the metaphoric picture. According to this classification, there are two types of visual metaphors: explicit and implicit. An explicit metaphor will include the product itself in a metaphorical illustration. On the contrary, an implicit metaphor will not include a product that may be displayed in a less visible place or be veiled (Chang, Wu, Lee, Chu, 2018). A Shell petrol advertisement from the 1930s is a good illustration of an implicit metaphor. The advertising tagline is "For the utmost horsepower". The cover depicts a stylised iron horse metaphorically characterising a vehicle, on the sides of which there is a harness in the form of canisters on which is written "Shell". Such fuel gives the "horse" incredible strength, and it soars from this power. The canister used in the form of a harness is a visual similarity to the usual attributes of a horse and, at the same time, a vehicle. An explicit metaphor is often used in advertising practice, in which the interpreter does not need to spend time searching for deeper meaning. For example, this kind of metaphor can be represented by an advertisement in which there is a group of people, most often a family in a friendly and happy atmosphere. Such advertisements ultimately suggest that happiness lies in the advertised product or necessarily includes it as a component through the use of visual codes, provoking familiar associations for consumers. It follows the fairly obvious conclusion that if such a product has already brought happiness to people on an advertising poster, then, accordingly, everyone has the opportunity to find it in the same way as they do. Using similarities in the analysis of metaphor, common semantic attributes of the significant and the signified can be found. Aside from that, commercial similarity attributes can serve as an incentive motivation to purchase for the customer.

Metonymy is a rhetorical figure of speech that plays an important role in the analysis of metaphor. Metonymy is a figure of speech in which one word is replaced by another selective or adjacent. Eco emphasises the close association between metaphor and metonymy. The author claims that any metaphor can be reduced to a chain of metonymic connections that make up the framework of the code, with the help of which the signified is correlated with the significant and serves as a support for any semantic field, as a field of possible meanings of a metaphor (Eco, 2005, p. 118). Consider an example of the analysis of the metaphor "visual noise", using one of the possible metonymic chains in which there is movement from one part of the metaphor "visual" to the other part "noise", on the assumption of Table 1.

Table 1

Explanation of the Metaphor “Visual Noise” Using Metonymic Series

Visual	Aural
Visible	Audible
Eye	Ear
Colour	Sound
Set of photons	Set of sounds
Chaotic photons	Chaotic sounds
Noise	

The table contains two metonymic series. Each column is a metonymic series of related notions related to different sensory organs. The first column is a series of related notions that reflect a semantic field, the differentiating feature or seme of which is the concept of “vision”. The last row of both columns is common, but only the second column, according to the literal expression, can contain a “noise” cell. The combination of columns, in this case, is possible using an analogy that allows us to establish a relationship between the “chaotic state” of the penultimate rows of both columns. As a result, we get the metaphor “visual noise” at the intersection of two semantic fields. The decomposition of a metaphor into this kind of series can indeed constitute an efficient tool for interpretation, but this is not the basis for its creation since only the first column without analysing the second column does not allow us to track the possibility and validity of finding the term “noise” in the semantic field “vision”. The metaphor creates a new semantic combination that can be analysed and explained using metonymic chains, which therefore can have a large number of variations due to the individual preferences of the interpreter.

2. Analogy as a Method of Generating and Analysing Metaphors

Analogy is one of the possible ways to create a metaphor which can be seen in the analysis conducted by Eco using the theory of interpretants. The author constructs Model Q (Model of Quillian), which is a set of nodes interconnected by various associative connections. Within the framework of this model, each sign is determined through interconnections with other signs that play the role of interpretants, each of which inversely can be a sign by itself. Eco puts to use the model to build a paradigmatic relationship system based on some code, which has the following form:

$$\begin{tikzcd}
A \arrow[d] & \text{vs.} & B \arrow[d] & \text{vs.} & C \arrow[d] & \text{vs.} & D \arrow[d] \\
k & & y & & z & & k
\end{tikzcd}$$

The horizontal lines form the paradigm of the sememe, and the verticals form the relationship between the sememe and the seme, or the semantic feature k (k is the semantic feature of A). If we denote A by k, then we can deal with a synecdoche or metonymy, since A and k are related concepts within the same semantic field. The seme k is inherent in the two sememes A and D. Therefore, by k we can, instead of A, put D, which will be a metaphor (Eco, 2005, pp. 136–137). This conclusion is nothing more than an example of the analogy of Aristotle, which finds application in the context of: "When the second word refers to the first in the same way as the fourth to the third, instead of the second you can put the fourth, and instead of the fourth, the second" (Aristotle, 1983, p. 669). Reformulating Aristotle's analogy into the model that Eco uses, we find two possible results of obtaining a metaphor, and not one, as Eco claims in his example.

1. A fundamental example of an analogy is the case of the existence of different semantic features in two different sememes, the re-setting of which allows us to find a metaphor.

$$\begin{tikzcd}
A \arrow[d] & \text{vs.} & B \arrow[d] \\
k & & y
\end{tikzcd}$$

Metaphor as a semiotic sign is not only inherent in the literary text, it can also be found in various semiotic texts. Consider the metaphor revealed in the architectural text, with its inherent codes to identify the principle of analogy. Reflecting on the anatomy of architecture, Sergey Kavtaradze observes that the use of metaphors, especially marine ones, is quite popular. In the church of the Holy Wisdom built at Constantinople (Istanbul) in the 6th century CE (532–537), the basilica consists of naves—ships, there are Anker—anchors that fix (anchor) metal rods, and these triangles were called sails. The dome on the sails is one of the most important elements of the alphabet of overlaps. If we consider the murals of the Christian church, these elements will surely include images of the evangelists—Matthew, Luke, Mark, and John. There are four of them and they support the church as well as the sails—the dome (Kavtaradze, 2015, p. 74). As may be inferred from examples, the analysis of the architectural text includes a set of different metaphors. To consider the principle of analogy, let us appeal to an example that takes an absolute form when the Anker refers to the nave, like an anchor to a ship. We represent such an analogy relation as a proportion in which: "Anker" / "nave" = "anchor" / "ship". In general terms, according to the formalisation of Eco, this will correspond to the expression “k” refers to “A”, as well as “y” to “B”, $k / A = y / B$, the outcome of this proportion will be the equality of the relations $k / y = A / B$, which leads to the conclusion: “Anker” / “anchor” = “nave” / “ship”. The final equality based on the proportion of analogy can indeed be the foundation for creating the metaphors presented above and translating them into architectural forms.

2. The second type of analogy is its special case, in which we have one semantic feature in two different sememes, the re-setting of which allows us to find a metaphor.

$$\begin{tikzcd}
A \arrow[d] & \text{vs.} & D \arrow[d] \\
k & & k
\end{tikzcd}$$

Such a connection between semes and sememes is a special case of explaining the metaphor by analogy because in proportion there will be the same element in a strictly established place. This type of connection will occur when “k” refers to “A” and also “k” to “D”, which ultimately leads to the expression $A = D$. Let us analyse the example of Eco, where the seme is a long white neck, sememes: a beautiful woman and a white swan, which accordingly gives the right to assert that a beautiful woman = a swan. The proportion of the analogy for analysing the metaphor will look like this: “long white neck” / “swan” = “long white neck” / “beautiful woman”. If we translate this statement into a proportion of analogy, we arrive at the following: $k/A = k/D \Rightarrow k\times D = k\times A \Rightarrow A=D$. As a result, analysing only the final expression $A = D$, it is necessary to understand that the addressee can find another seme for the interpretation that formed the equality (for example, a woman is called a swan because of grace and beauty), or else completely refute this kind of equality, saying that the long neck does not give beauty and resemblance to a swan. A special case of analogy can be applied only in the established order when semes and sememes are at the same level in proportion. If the order is not followed, as, for example, in the case $a / b = c / a$, where the element “a” is also in both parts of the proportion, this leads to the expression $a \times a = c \times b$, which in the analysis of the metaphor is devoid of truth, since two identical elements (a, a) do not create metaphors.

The aforecited model is a method of analysing metaphors using the proportion of analogy. Such a model can also be used to analyse non-metaphorical rhetorical figures of speech, which will be revealed on the basis of equality of the seme, including the same seme, or inequality, including different semes. This is evident in an example with the help of which it is possible to establish this kind of equality and inequality on the basis of the analysis of a musical work as a semiotic text proposed by Raymond Monelle in the article “Music and Semantics”. This is illustrated in the work undertaken by the outline of the analysis of

Wagner's musical piece "Tristan-Prelude", which the author proposes using Table 2 (Monelle, 1995, pp. 105–107).

Table 2

Semantic Analysis of a Musical Piece

	Motives and meanings	
Sememes	Semes	Leitmotiv
α	W + C	Confession of love; grief, sorrow
β	C + F	Desire
δ	D + T + W + F	The glance
ε	D + C+ T + W + F	The love-philtre
ζ	D + T + F	The magic casket
η	W (+ E)	Death

Note. The content of the table comes from the work of Monelle (1995, p. 100).

A musical text has a complex structure and many different elements that can be semes, such as a single note or rhythm. Sememes are larger phrases or sentences in musical formations. In the presented scheme, we analyse two sememes α and β. Raymond Monelle, being a music expert, comparing two sememes that are heterogeneous in nature, finds in them a common element C, which is a chromatic scale—a way to organise a series of musical notes in height. The difference leading to a musical debate is that the sememe α begins with the chord W, and the sememe β ends with the chord C, which allows for the distinguishing of confession of love from desire. In a similar manner, the constituent parts of sememes can be analysed, however, the fact of the existence of different and similar semes in one sememe is the basis of a two-sided analysis using the Model Q. To apply the model, we rewrite the part of the circuit of Figure 1 containing the sememes α and β in the following form:

$$\begin{tikzcd}
\alpha \arrow[d] & \text{vs.} & \beta \arrow[d] & \alpha \arrow[d] & \text{vs.} & \beta \arrow[d] \\
\mathrm{C} & & \mathrm{C} & \mathrm{W} & & \mathrm{F}
\end{tikzcd}$$

The application of the model for the analysis of the sememes α and β reveals both of the previously considered possible cases when there is one similar seme C and two different semes W and F. Initially, the sememes have the following form: α=C+ W and β=C+ F, this suggests that α and β have semes, with the help of which one can conclude both equality and inequality between them. Equality, which will be concluded by analysing the proportion of analogy, can reflect not only the metaphorical relationship between the sememes but also represent other figures of speech such as metonymy, comparison, similarity, conformation, etc. In this case, we can conclude that the sememes α and β are equal, based on the comparison with the help of the seme C, or the sememe α is more emotionally calm and gentle than β based on the comparison of the chords W and F. However, both schemes, with one common seme or with different ones, can be a prerequisite for creating a metaphor. As a result, the interpreter decides himself on account of which seme he concludes the analogy between the sememes and whether this analogy is generally a source of metaphor formation. The cases reported here illustrate that we can really deal with a metaphor using an analysis of analogies, however, this kind of attitude, after all, does not always form a metaphor. As well as the fact that the principle of analogy, full or special, can be one of the methods for generating metaphors, it can also be a tool for its analysis, when applied in the reverse order.

We can return to the example of the “visual noise” metaphor to show how the metonymic series can be part of the analysis of analogy. The initial link of the metonymic chain will be at the same time an integral part of the metaphor and one of the semes. Each metonymy in the chain will be nothing more than a possible seme of the sememe. The first column is a chain of k-metonyms of the “Visual” sememe, the second column is the y-metonymy of the “Aural” sememe. It should be noted that the same metonymy can be part of different sememes. I present this statement in the form of Table 3.

Table 3

Metonymic Series

	A		B
	Visual		Aural
k_1	Visible	y_1	Audible
k_2	Eye	y_2	Ear
k_3	Colour	y_3	Sound
k_4	Set of photons	y_4	Set of sounds
k_5	Chaotic photons	y_5	Chaotic sounds
		y_6	Noise

The decomposition of the sememes into this kind of metonymic series, which is a set of semes, is an important point in the analysis of metaphor, which can be identified based on what seme (links of the metonymic chain) the following analogy is drawn. Considering the sememes A and B, it should be noted that the following prerequisites for constructing the analogy proportion are the most preferred semes: k5 are chaotic photons and the last cells y5 and y6 of the sememe B, which can be combined into one, since a set of chaotic sounds is equal to noise. Given this, an analogy will be constructed based on the A-visual, B-aural sememes, semes k—chaotic photons—and y—noise.

$$\begin{tikzcd}
A \arrow[d] & \text{vs.} & B \arrow[d] \\
k & & y
\end{tikzcd}$$

This case has shown that, based on the analysis of metonymic series, the most significant predicates (semes) of the metaphor parts (sememes) are revealed, which can be used to construct the analogy proportion for subsequent analysis.

Thus, three cases of analogy can be distinguished as a method of metaphor generation. The analogy can be represented as proportions:

1. $a / k = b / m$—the case when the equality of relations of objects “a” and “b” with the semantic attributes of their semantic fields “k” and “m” is established;
2. $a / a_1 = b / b_1$—the case when the equality of relations of objects “a” and “b” with other objects of their semantic fields “a_1” “b_1” is established;
3. $a / x = b / x$ or $x / a = x / b$—the case when the equality of relations of objects “a” and “b” with the same parameter characteristic (x) of both objects is established.

3. Metaphorical Transfer as a Method of Generating and Analysing Metaphors

Fundamental in the process of metaphorisation is the concept of metaphorical transfer, which in the framework of semiotics is a deeper and more complex process than in the traditional theory of metaphor. One of the earliest examples of the mention of such a process is associated with the name of Aristotle and his work “Poetics”: “A metaphor is an unusual name transferred from genus to species, or from species to genus, or from species to species, or by analogy” (Aristotle, 1983, p. 669). Glazunova, studying the logic of metaphorical transformations, emphasises that “metaphorical transfer is a transfer of meaning from one object to another” (Glazunova, 2000, pp. 177–178). The reason why this kind of transfer creates metaphorical relations between different objects could be found in the cognitive view on the phenomenon of metaphor. Lakoff and Johnson define metaphor as a way of thinking and understanding one thing as and in terms of another thing (Lakoff, Johnson, 2008, p. 62) Metaphorical transfer is carried out to comprehend the object of one semantic field with the help of another object of another semantic field. As noted earlier, the process of signifying occurs mutually, each object influences the meaning of the other. When considering this process within semiotic studies, it is necessary to note its contiguity with the concept of Pierce’s parallelism. Parallelism and metaphorical transfer allow two different objects to be in the same semantic field and participate in the process of mutual denotation. The parallelism is an important vehicle for semantic innovation as it creates new possibilities, new combinations, and new semantic couplings (Sørensen, Thellefsen, 2006, pp. 207–210). For example, through the metaphorical transfer in the established metaphor “time flows”, “time” as an object takes some meanings of “fluidity”, and the term “flow” takes over some connotations of “time”, which ultimately makes it possible to intersect the metonymic series of both terms. In addition to the fact that transfer occurs between the meanings of various objects, in semiotics it can also occur between different semiotic systems, since a metaphor, as a semiotic sign, is not only inherent in literary texts and the field of rhetoric, here it acquires a place in other semiotic systems. The rhetoric within the framework of semiotics represents the transfer into one semiotic scope of the structural principles of another (Lotman, 2002, p. 201). From these considerations in semiotic texts we can be concerned with metaphorical transference within one semiotic system and between different ones. Many semiotic metaphors retain their verbal nature, so the transfer can take place either at the border “verbal text / another semiotic text” or vice versa. Examples of transference of “verbal text / another semiotic text” can be found at the junction of the arts, which the surrealistic work of Salvador Dali demonstrates. One such example of a metaphorical transfer of the “word / sculpture” type is his well-known work “Venus of Milos with Drawers”. The master’s sculpture reveals several metaphorical associative expressions such as an eternal search for something important, to rummage, to intrude on someone’s feelings, to dig into someone’s soul, to search for meaning, self-chastise, being in one’s head. According to this example, we shall analyse the metaphorical transfer within the “get inside someone’s soul” metaphor as a verbal metaphor and as a metaphor for Dali’s sculpture. For this purpose, we shall construct Table 4, which will characterise the interpreter’s metonymic series concerning the “to get inside” and “soul” objects and will be filled in according to the degree of correlation between the rows and columns in each cell on a probability scale of [0, 1]. The probability scale is a subjective numerical value that the interpreter ascribes to each cell as the most comprehensible and theoretically possible merger of two different terms and their semantic fields.

Table 4

Metonymic Series of Interpretation of the Metaphor “Get Inside Someone’s Soul”

	Soul	Heart	Inner world
To get inside	1	0,5	0,6
Interfere	0,6	0,4	0,7
Break in	0,8	0,6	0,7

The table shows that the metaphor “get inside someone’s soul” out of context gives the interpreter several associations that are embodied in the presented metonymic series. Such causes can include an infinite number of rows and columns, depending on the preferences of the interpreter. The horizontal and vertical axes of the table show the degree of influence of the term “to get inside” and the term “soul” between themselves as parts of a metaphor. The first numerical column is a reflection of the relationship of the "soul" with the metonymic series of the term “to get inside”, the first row, respectively, is the other way around. Based on the numerical data, it is possible to analyse in what relation the parts of the metaphor influence each other on the basis of the arithmetic average of the first columns and the row, which will be an indicator of how much this or that part of the metaphor at the transfer can belong to another metonymic series. In this example, such an indicator of the first row is 0.7, which means that the term “to get inside” with a high probability may belong to the metonymic series of the term “soul”. The indicator for the first column is 0.8, which indicates that the term “soul”, although not by much, is still more acceptable for being in the metonymic series in which the term “to get inside” is placed. However, both terms have a rather high influence on each other in the process of signification formed by metaphorical transfer, which allows the terms to be reflected in each other’s metonymic series based on the intersection of their semantic fields.

In the following, we turn to the analysis of the second example of the “get inside someone’s soul” metaphor based on the analysis of Dali’s sculpture “Venus of Milos with drawers”, as shown in Table 5, constructed on the same principle as Table 4.

Table 5

A Metonymic Interpretation Series of the Metaphor “Get Inside Someone’s Soul” With the Example of a Sculpture by Dali “Venus of Milos With Drawers”

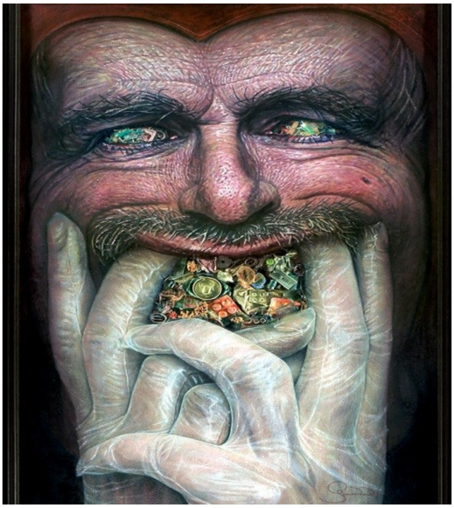
	Soul	Drawer	Emptiness
To get inside	1	1	0,7
Look for	0,9	0,6	0,7
Open	1	1	0,7

In this example, it is necessary to notice the difference in the metonymic series of the metaphor compared to the previous one, due to the specific context that the sculpture creates. Here, the index of the first line is 0.9, which indicates that the term “to get inside” with a high probability may belong to the metonymic series of the term “soul”. The indicator for the first column is close to unity and amounts to 0.97, which indicates that the term “soul” is not just more acceptable in the metonymic row in which the term “to get inside” is placed, but that these terms are with high probability parts of the same metonymic chain. Compared with the previous example, in the context created by the sculpture, the metaphor finds a deeper understanding based on numerical indicators. Having used visual codes and context, the terms that make up the metaphor find the possibility of a high degree of belonging to each other. An analysis of the metaphorical transfer between two different semiotic systems clearly shows how the same metaphor, according to its peculiarity of openness and ambiguity, can have an infinite number of interpretations in different cultures and among different interpreters.

A different kind of metaphorical transfer can be the reverse of the previous process, when the verbal text is not the previous one, for example, it can be embodied in the “taste/word” scheme. Taste codes open up a wide range of connotations and synesthesia, forming such metaphorical transfers, such as “sweet life”, “the bitter truth”, etc. (Eco, 2004, p. 500). Aside from that, the result of such a transfer can be a metaphor already presented in the framework of another semiotic system. So, the metaphor “the bitter truth”, for example, is embodied in the works of the modern Polish artist Krzysztof Grzondziel. The author depicts various realities of the “truth” of the modern world such as terrorism, insensibility, deception, the devastation of people, the destruction of the environment, and much more, which he, in turn, exposes in his artistic metaphors. One example of a metaphor that reflects the “the bitter truth” of modernity is contained in the artist’s self-portrait, shown in Figure 1. The painting depicts a man, inside of whom there are only shreds, scraps, rotten garbage, the remains of some things, packaging of well-known brands, and pills. Could it be all that a person managed to accumulate in his life and fill himself with? The “bitterness" of this work lies in the fact that of the whole set of these things there is really nothing to carry away, and there is no free space to fill with something else valuable.

Figure 1

Self-Portrait by Krzysztof Grzondziel



The metaphor “the bitter truth” in the final instance can have completely different connotations and interpretations from the original verbal form, representing a whole series of works, semiotic texts that will characterise it based on a chain of “taste / word / image” transferences. Thus, with the help of transference, the metaphor stands to gain “life” in various semiotic texts, which makes it possible for the terms that it includes to acquire a place in various metonymic series, expanding the boundaries of possible meanings both individually and within the metaphor.

The rarest type of metaphorical transfer should be called one that does not include a verbal text. An example would be the “wordless” metaphor of architectural texts, as in the analysis of analogy discussed earlier. In addition, it should be emphasised that the metaphor in architecture not only forms an image, but also affects the technology and idea of the invention. Such a metaphor is devoid of verbalisation; it gives it an unusual form, an individual perception and interpretation. Thus, metaphorical transfer or parallelism between two different terms and their metonymic series enables the formation of metaphors. The same metaphor can be generated by transferring between different metonymic series, depending on the semiotic text in which it is embodied, which makes it unique in interpretation.

4. Metaphor as a Message in the Communication Process. Encoding and Decoding

The next important aspect of the analysis of metaphor as a semiotic sign is the consideration of the features of its embodiment in the communication process. The standard communication model includes: a sender, an addressee (recipient), a message. The message, in turn, is interpreted using certain codes. When operated on a metaphor, it is clear that it is a message in itself, the sender can be any kind of semiotic text in which this message will be encrypted and the addressee is any interpreter that encounters the text of the sender. Eco emphasises that various codes and subcodes may participate in such a process, depending on sociocultural circumstances. Such codes may differ between the addressee and the sender, since the addressee may put forward their initial presuppositions and explanatory hypotheses of abduction (Eco, 2005, p. 14). Some types of metaphors oblige the addressee to have a certain arsenal of subcodes that will be shared with the sender. This condition is necessary to realise the understanding of metaphor on both sides. This can be most clearly shown by the example of philosophical metaphors of the form: “Russell’s teapot”, “Occam’s razor”, “Diogenes barrel”, etc. Expressions such as Kant’s “thing in itself” or Nietzsche’s “superman” are not artificial “exotic” words, but are terms that give rise to a new discourse and reasoning (Tulchinskii, 2019, p. 66). I shall try to show that these are metaphors of a special kind, which at first glance represent some “exotic constructions” that are explained by their creators, which is more like the name of ideas, a certain “exotic” slogan. The names used in the metaphors above are the most difficult part because they constrain the addressee to be familiar with the text and explanations of the author. The rest of the words—like “teapot”, “razor” “barrel”—are exactly metaphorical. Do they need names, their authors, to understand the meaning attributed to them? In these exclusively philosophical metaphors, the name implies a link to an explanation and acquaintance with the author. The “teapot” acquired a new meaning with the help of “Russell,” and “Russell,” as a philosopher, acquired a new meaning for his name through the “teapot”. However, knowledge of the name does not provide a basis for understanding the metaphor, it is important to note the need for full knowledge of the author’s description of the meanings. The interpreter may use just such an explanation, or may have his own, but still based on the text of the owner of the Name. This suggests that some types of metaphors require that the message-expression, as a source of information, and message-content, as interpreted text, have at least one common subcode for understanding.

In interpreting the metaphor as a message, a significant role is played by various codes that are used by the sender and the addressee. The difference in codes is not only a feature of the perception of each individual, it is also formed per the form of the content of a particular metaphor. Thus, the foundations of parallelism can be found in such a property of the code as rule-governed creativity. Eco observes that the code, using the well-known elements of culture, allows one to generate assessments about facts, manipulating the significant to correlate them with the new signified (Eco, 2005, p. 118). Consider the example of the metaphor “drown in the eyes” in two different forms that represent it. Table 6 shows one of the options for analysing the metaphor by the addressee based on the “Eye to Eye” drawing by Edward Munch. The first column is a metonymic series that expresses the peculiarity of understanding the element “drown”, as one of the parts of the analysed metaphor. The first line is a mapping of a series of interpretations that describe another part of the metaphor, the element “eyes”.

Table 6

Analysis of the Metaphor “Drown in the Eyes” With the Example of the Work of Munch “Eye to Eye”

	Eyes	Whirlpool	Darkness
Drown	The gaze of both characters	Touch covered in darkness	Dark gloomy background
Wreck	Blurred image of a girl’s face	Sad, saddened look of a couple	Shadow filling the girl’s face
Fear	The expression of the dark eyes of the girl	Dividing tree in the middle	The pale face of the many in the dark

The first row and column are the formed verbal interpretation of codes presented in the form of metonymic series. The codes themselves are the contents of the table. The filling of the table can vary, for instance, “a dark gloomy background” for some can be a connotation of “fear of the darkness,” and not a characteristic of the expression “drown in the darkness” as in the example. Codes and their interpretations by another addressee may differ from the one presented. What is important to us is the course of analysis, which consists in understanding the peculiarities of the metaphor within a certain form, using the codes of the semiotic text and building the appropriate metonymic series.

In the following, we consider Table 7, built on the same principle as the previous one, which presents an analysis of the “drown in the eyes” metaphor based on the poem by Rozhdestvensky “May I sink in your eyes?”. A fragment of which is presented below.

May I sink in your eyes?
Because sinking in your eyes is happiness.
I will come to you and say: "Hello,
I love you". It is complicated…
No, it is not complicated, it is hard
It is very hard to love, do you believe it?
If I come to the edge of the cliff
And fall down, will you come in time to catch me?
And if I am away, will you write to me?
I want to be with you for a long
For a very long time…

Table 7

Analysis of the Metaphor "Drown in the Eyes" With the Example of the Poem by Rozhdestvensky “May I Sink in Your Eyes?”

	Eyes	Happiness	Love
Drown	If I come to the edge of the cliff	Sinking in your eyes is happiness	I want to be with you for a long
Speak	Tell me with your eyes, do you love me?	I am afraid to get your answer, you know… Tell me, but tell me silently	I will come to you and say: “Hello, I love you”
Fear	Not to blame me with your look	Not to take me to the deep waters	May I love you? Even if I must not, I will!

When comparing the data of two tables revealing the same metaphor, first of all, it is necessary to emphasise the general background, as the context prevailing at the addressee. In the first case, this is the gloomy appearance of two people who are not indifferent to each other, who, however, are in the darkness of their own eyes and the world around them. Codes and their interpretation encounter sadness and regret, they do not reveal the hope of the possibility of salvation in the eyes of the opposite. The second example, although contained in verbal form, gives a sense of light tones. The poem asks many questions for which there are no answers, but still, it feels that the messenger is shrouded in warm feelings, and perhaps does not need any answers, since the eyes of a loved one have already given him joy for which he is able to approach a steep cliff, sacrifice everything in order to "drown" again, in order to be saved.

The analysis of codes and their interpretations in various forms opens up innumerable options for understanding meaningful metaphors. Decoding, with which the metonymic series is built, allows the addressee to further analyse using a probability scale to calculate the most significant codes depending on the embodiment of the metaphor and the cultural environment of the interpreters.

5. Conclusion

The metaphor, combining two or more objects together, makes it possible to comprehend one in the other, eventually forming the ambiguity of the possible outcomes of the analysis. Such a plurality of interpretations should be called openness. Considering the metaphor as a sign that we use when referring to our natural or cultural environment, it is necessary to emphasise its dependence on how language or other sign systems define things. Metaphors are produced solely on the basis of a rich cultural framework, on the basis, that is, of a universe of content that is already organised into networks of interpretants (Eco, 1984, p. 127). The metaphor, as part of various semiotic systems, allows for multiple analysis of the foundations of its creation and subsequent interpretations. One of the more significant findings to emerge from this study is that methods of analogy and transfer as the main forms of metaphor generation and as methods of its analysis open up new facets of understanding and studying this phenomenon. This study has shown that the mechanism of generating the same metaphor within the framework of semiotics can vary and have numerous forms for embodiment, which complicates and deepens the process of its analysis. To search for the foundations of a metaphor, which can be different for both the interpreter and the creator of the text in which the metaphor is used, and among interpreters in general, it is necessary to focus on an important part of any type of analysis—the detection of metonymic series. Metonymic series reflect a list of associations connected with terms included in metaphorical relationships based on an analysis of codes, context, or other cultural or subjective considerations. Further, from the perspective of the possible methods of creating a metaphor, the interpreter can apply analysis based on finding the principles of similarity, analogy, or transfer.

The proposed methods are aimed at achieving a comprehensive analysis of the possible meanings and foundations of the metaphor for a multifaceted understanding of the object under study. The research has also shown that such methods of analysis are expedient to use both for those who embody the metaphor in some text, and for the interpreter. The formulation of offered methods of metaphor generation and analysis as the purpose and novelty of the paper allows a description of metaphorical relations between different objects and the openness of the phenomenon.

REFERENCES

Aristotle. (1983). *Поэтика* [Poetics] (Complete Works in 4 Volumes, Vol. 4). Moscow: Mysl.

Chang, C.-T., Wu, Y.-C., Lee, Y.-K., Chu, X.-Y. (2018). Right Metaphor, Right Place: Choosing a Visual Metaphor Based on Product Type and Consumer Differences. *International Journal of Advertising*, *37*(2), 309–336.

Eco, U. (1984). *Semiotics and the Philosophy of Language*. Bloomington and Indianapolis: Indiana University Press.

Eco, U. (2004). *Отсутствующая структура. Введение в семиологию* [The Missing Structure. Introduction to Semiology]. St. Petersburg: Symposium.

Eco, U. (2005). *Роль читателя. Исследования по семиотике текста* [The Role of the Reader: Explorations in the Semiotics of Texts]. St. Petersburg: Symposium.

Glazunova, O. (2000). *Логика метафорических преобразований* [The Logic of Metaphorical Transformations]. St. Petersburg: Faculty of Philology, State University.

Kavtaradze, S. (2015). *Анатомия архитектуры. Семь книг о логике, форме и смысле* [Anatomy of Architecture. Seven Books on Logic, Form and Meaning]. Moscow: Publishing House of the Higher School of Economics.

Lakoff, G., Johnson, M. (2008). *Метафоры, которыми мы живем* [Metaphors We Live By]. Moscow: LKI publishing house.

Lotman, Y. (2002). *Статьи по семиотике культуры и искусства* [Articles on Semiotics of Culture and Art]. St. Petersburg: Academic Project.

Morris, C. (2001). *Основания теории знаков* [The Grounds of the Theory of Signs]. Ekaterinburg: Academic Project, Delovaya kniga.

Peirce, C. S. (2000). *Избранные философские произведения* [Selected Philosophical Works]. Moscow: Logos.

Rosengren, M. (2019). О созидании, пещерном искусстве и восприятии: доксологический подход [On Creation, Cave Art and Perception: a Doxological Approach]. *Journal Voprosy Filosofii*, *8*, pp. 80–93.

Sørensen, B. (2011). The Concept of Metaphor According to the Philosophers C. S. Peirce and U. Eco—a Tentative Comparison. *Signs—International Journal of Semiotics*, *5*, 141–176.

Sørensen, B., Thellefsen, T. (2006). Metaphor, Concept Formation, and Esthetic Semeiosis in a Peircean Perspective. *Semiotica*, *2006*(161), 199–212.

Monelle, R. (1995). Music and Semantics. In E. Tarasti (Ed.), *Musical Signification: Essays in the Semiotic Theory and Analysis of Music* (pp. 91–108). Berlin: De Gruyter Mouton.

Tulchinskii, G. (2019). Философия как проектирование новых смыслов [Philosophy as Design of New Meanings]. *Journal Voprosy Filosofii*, *7*, pp. 64–68.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXV, nr 1 (2021), s. 73–94
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxv1.05

Article

PIOTR KOZAK [*]

THE ANALOG-DIGITAL DISTINCTION FAILS TO EXPLAIN THE PERCEPTION-THOUGHT DISTINCTION: AN ALTERNATIVE ACCOUNT OF THE FORMAT OF MENTAL REPRESENTATION[1]

SUMMARY: The format of mental representation is the way information is organized in the mind. The discussion surrounding the format of representation addresses the problem of what representational primitives are and the rules of information processing.

In philosophy, the discussion is dominated by the distinction between analog and digital representational systems. It is thought that this distinction can bring us closer to an understanding of the nature of perceptual and discursive representations.

I argue that the analog-digital distinction cannot meet that expectation. The analog-digital distinction is neither sufficient nor necessary to explain the distinction between perceptual and discursive representations (and perception and thinking, respectively). I propose an alternative interpretation of the concept of representational format which provides us a better understanding of the difference between iconic and discursive representations. I explain the differences between formats of representations in terms of differences in information processing. I demonstrate, how this alternative interpretation of the concept of the representational format can explain the constraints put on the contents of representational systems.

KEYWORDS: mental representation format, analog, digital, perception, thought, iconic representations, discoursive representations.

[*] University of Bialystok, Institute of Philosophy. E-mail: piotr.kozak1@gmail.com. ORCID: 0000-0001-9734-4640.

[1] This work was supported by the research grant “What is Thinking with Images?”, SONATA 10, granted by the National Science Centre, Poland, based on decision No. 2015/19/D/HS1/02426.

A mental representation format is the way information is organized in the mind.[2] A discussion of mental representation formats addresses the question of how the information in our mind is stored and processed. It concerns the structure of representations interpreted as a set of representational primitives and combinatorial principles. Thus, to describe a representational format, one has to describe the structure of representation, that is, one has to explain what the primitive elements are and the possible operations that can be carried out with them.

There are two philosophical traditions of thinking about the format of representation. First, following Goodman (1976), we distinguish between analog and digital systems of representation. Second, Goodman's distinction between two formats of representational systems has been adopted in the philosophy of mind as the basis for the distinction between perceptual and discursive representations. Following Dretske (1981), it is argued (Peacocke, 1989) that perception consists of iconic representations and is analog in format. In contrast, beliefs and thoughts are discursive representations and are encoded in digital format.[3] In this paper, I will use the terms "iconic" and "discursive representation" as referring to representations that describe perceptual and discursive mental phenomena, respectively.

Let me give two examples of the discussion surrounding representational format. First, the early stage of the so-called imagery debate (from the 70s to the beginning of the 90s) was mostly devoted to issues surrounding the way our minds encode mental images. On the one side, pictorialists (e.g., Kosslyn, 1980) have held that the format of mental images is perceptual-like and analog. On the other side, descriptionalists (e.g., Pylyshyn, 1973, 1981) have argued that mental images are formed out of structured descriptions that are digital in format. Most descriptionalists have argued that mental images are epiphenomena of some internal discursive, language-like processes.

Second, the debate on representational format underpins philosophical discussions on the nature of perceptual representations. On the one hand, some philosophers, most notably Sellars (1997) and McDowell (1996), have held that perceptual representations are propositional. On the other, there are philosophers, such as Crane (2009), Dretske (1969), and Travis (2013), who have argued that perceptual representations are distinct in kind from discursive representations. For those who deny that perceptual representations are propositional, the question arises what kind of representations they can be. The most common answer is that perception has an iconic structure. It consists of iconic elements and relations between these elements. The "atoms" of perception are iconic representations that are most often described as having an analog nature and as being deprived of canonical decomposition. Interactions between these elements are based on causal

[2] The concept of the format of mental representation is different from the concept of the vehicle of representation. Information can be stored in the same format in different types of vehicles. Using a computer metaphor, the same .jpg file format can be stored on both magnetic vehicles, as in the case of a floppy disk, and optical vehicles, as in the case of a CD.

[3] That does not mean that there are no intermediate representations, such as maps and pictographs, see Casati and Giardino (2013).

relations between iconic representations. In contrast, interactions between discursive representations are based on logical transitions (e.g., Matthen, 2005).

The difference between analog and digital formats of representation is intuitive but conceptually blurred. According to Goodman (1976), to be an analog representational system means to be both a syntactically and semantically dense representational system. An example of a dense representational system is an old-fashioned clock that represents time continuously, unlike a digital clock that represent time discreetly. Moreover, an analog representational system is relatively replete. A representational system is relatively replete if, in comparison with other systems, many of its members' features are relevant to determining what they represent. The system of old-fashioned analog clocks is not replete, since only the position of the clock's hands matter. In comparison, in the case of images, such features as colour, shape, and size are relevant. However, for reasons that will not be covered here (e.g., Kulvicki, 2006), it is doubtful whether Goodman succeeded in adequately explaining analog and digital formats of representation.

In the last 50 years, the distinction has been variously interpreted and explicated (e.g., Fodor, Pylyshyn, 1981; Haugeland, 1998; Lewis, 1971; McGinn, 1989). Across those approaches, digital representations are generally understood to be discrete entities. Numerals provide a good example. "0" and "1" are discrete because they indicate distinct and separable entities. For every representational token, it is clear which type it instantiates. In contrast, analog representations do not admit definite type-identity. For example, the colour value of a given colour patch is measured on a continuous rather than a discrete scale (Dretske, 1981). There is always room for the question of whether a given colour patch is more blue-like or dark blue-like. In contrast, whereas there is no room for the question of what number is represented by "0".

Iconic representations are believed to be analog structures. For instance, there is no way to determine a discrete point where a blue colour patch ends and a dark blue one begins. The structure of discursive representations is believed to be based on digital operations. For instance, if one believes that the king is dead, one thinks about the king, and not more or less about the king, ascribing the property of being dead and not being more or less dead to the king. In contrast, a mental image of the dead king more or less resembles the king being dead.

However, this understanding of the analog-digital distinction is far from being clear (e.g., Lorenzo, Rubiera, 2019; Maley, 2011). A colour patch can be represented in analog format but can be represented digitally as well, namely, as a set of colour values in the RGB colour model. Musical notes C and C# are discrete when playing piano, but there is a continuum of notes between C and C# when playing a violin. A film frame is discrete, but the events depicted with the help of film frames are indiscrete.

An alternative way to interpret the analog-digital distinction can be put in terms of constraints that the representational system puts on representational content. It can be illustrated with notational systems in mathematics. Although the same mathematical magnitudes can be recorded in different notational systems, such as Roman or Arabic numeral systems, these systems are not computationally equivalent. For example, it is more efficient to carry out a calculation with large numbers in Arabic than the Roman numeral system. Analogously, it may be more efficient to represent the values of a linear function with the help of a graph than with the help of a numeral matrix.

Thus, the difference between representational systems can be interpreted in terms of being capable of expressing different kinds of content. That means that different representational systems put constraints on the content a representational system can carry and the range of possible transitions between different contents. So, for example, an analog representation can represent the value of a magnitude but not an integer (Beck, 2015), iconic representations cannot be used to represent a negation (Crane, 2009), etc.

Yet these two interpretations are linked, for if one wants to explain why some representational systems put constraints on representational content, then one has to describe the features of the representational structure. For instance, one can explain why the Arabic numeral system is preferred over the Roman numeral system by pointing out the fact that the Roman numeral system does not have the concept of zero. Analogously, iconic and discursive representational systems put constraints on their representational contents. A theory of representational format should explain where these constraints come from.

To put it more generally, the question is what should the concept of the format of mental representation explain and how it can do that. I claim that the problem with the analog-digital distinction is not that it is not clear. Even if it were clear, it would still be doubtful whether it could explain what it should explain, namely, the difference between thoughts and perceptions.

In this paper, I propose an alternative interpretation of the representational format. I claim that it provides us a better understanding of what the difference between iconic and discursive representations is. In the next section, I show what any theory of representational format should be able to explain. I demonstrate how to interpret the difference between iconic and discursive representations. I claim that discursive representations can meet the requirements of the so-called Generality Constraint, while iconic representations cannot. Next, I explain how one can understand the difference between iconic and discursive formats of representation. I put it in terms of differences of information processing in cognitive systems. Last but not least, I demonstrate how the alternative interpretation of the concept of a representation format can explain the constraints put on the contents of representational systems.

1. Generality Constraint

One of the distinctive features of discursive representations is that they are systematically structured. Entertaining a thought of one kind entails a capacity to entertain a thought of another kind. For instance, entertaining the thought that John is happy and that Mary is sad is systematically connected with the cognitive ability to entertain the thought that Mary is happy and that John is sad. Having the thought that John is happy entails a capacity to think that someone is happy. In Evans' words (1982, p. 104), "if a subject can be credited with the thought that *a* is *F*, then he must have the conceptual resources for entertaining the thought that *a* is *G*, for every property of being *G* of which he has a conception".

The same rule applies to inferences (e.g., Fodor, Pylyshyn, 1988). If I think that it is dark and cold and raining, I can infer that it is cold and raining; for from *P* & *Q* I can infer that *P* (or *Q*). By the same token, I must be able to infer from it is cold and raining that it is raining. If I am unable to do so, I do not know what inference is.

Thus, discursive representations are systematically co-related (e.g., Heck, 2000; Peacocke, 1992), which means that they are systematic in nature. Evans (1982) calls this requirement the Generality Constraint.

To meet this requirement, discursive representations have to consist of recombinable constituents that can build more complex structures. It means that discursive representations are compositional in nature. Compositionality of discursive representations means, first, that the meaning of complex structures is determined by the meaning of their constituents. The constituents of discursive representations are parts of the representations that are canonically distinguishable, for not every partition of the representation makes sense. The idea is that canonically decomposed parts are syntactically and semantically meaningful units. The thought that John loves Mary can be decomposed into John loves and Mary, but not into John…Mary (e.g., Fodor, 2008).

Second, the meaning of complex discursive representations must come from the meaning of their canonically distinguishable parts together with the rules of composition, for not all combinations are allowed. The recombination of the parts must be meaningful. John loves Mary can be recombined into Mary loves John, but not into John Mary loves.[4] These rules are recursive. If I have a thought that John loves his mother, I must be capable of having the thought that John loves his mother's mother, etc. Putting it together, discursive representations have a recursive syntax that combines canonically distinguishable parts according to combinatorial rules (e.g., Pagin, Westerståhl, 2010).

Language seems to be systematic and compositional. It has syntax and distinguishable syntactic and semantic parts. Thus, one may infer that discursive representations are language-like representations (e.g., Devitt, 2006). In contrast, iconic representations lack systematicity and compositionality, and therefore they do not have the metaphysical properties we are looking to ascribe to discursive representations.

[4] According to Evans (1982), systematicity is constrained by semantic conditions of appropriateness. For instance, thinking that JOHN FELL INTO THE LAKE need not entail a capacity to think that THE LAKE FELL INTO JOHN. However, even if a well-formed string of thoughts is a semantical absurdity, it does not mean that it cannot express thoughts. For one thing, we can entertain absurd thoughts. For another, an absurd but well-formed string of thoughts can be the basis of inferences in logic. See, e.g., Camp (2004).

Iconic representations are neither systematic nor compositional, for they lack syntactic structure.[5] According to Fodor, they lack canonical decomposition. According to Frege, they lack logical form. Therefore, iconic representations do not meet the Generality Constraint.

Fodor's argument (2007; 2008) from lack of canonical decomposition takes the form of the so-called Picture Principle. According to the Picture Principle, iconic representations can be distinguished topologically: although pictures have interpretable parts, they lack canonical decomposition. It means, loosely, that we can cut up a picture however we like, and each picture-part will represent a relevant part of the represented object. Thus, every part of the representation represents some part of the scene represented by the whole representation (e.g., Green, Quilty-Dunn, 2017; Quilty-Dunn, 2016; 2020; Sober, 1976). In contrast, discursive representations have a canonical decomposition, which means that they cannot be cut into pieces however we like. Discursive representations have constituent parts. For instance, the content of the proposition snow is white can be decomposed into the parts snow and is white, but not into snow…white, which means that the expression "snow…white" does not possess independent semantical value. Thus, although iconic representations can be decomposed, they cannot be canonically decomposed. However, if they can be composed and decomposed however one wants, then they lack syntactical structure.

Frege's argument from lack of logical form is based on the observation that icons are unable to express logical relations. For logical relations to hold, the elements of the relation have to possess a logical form, i.e., syntactically fixed structure, such as a set of logical constants and variables, with determined transformational rules that preserve the logical values of its components. If A implies B, then basing on transformational rules it is possible to transform the truth of the first into the truth of the second. Propositional logic shows how the truth of complex propositions depends on the truth of simple ones. Truth-functions operate on propositions that can be negated, disjoined, and conjoined; they can imply one another or be equivalent. One of the main reasons for talking about propositions at all is that they explain how things can stand in these logical relations.

Iconic representations cannot express logical relations, for they lack logical form. There are no truth-preserving transformation rules for imagistic representation. There is no pictorial negation (Crane, 2009; Sainsbury, 2005), conjunction, or disjunction (Heck, 2007); images cannot express implications or quantifications (Frege, 1984), etc.

Frege's argument implies that a clear line between iconic and discursive representations can be drawn. Discursive representations are interpretable in logical

[5] That does not mean that they lack construction rules. They are obviously rule-governed. That is why if one understands how to interpret one Venn-diagram, then one understands how to understand another. However, a representational system can have construction rules without syntactic structure. I can construct a triangle according to the rules of construction, but that does not mean that triangles have a syntax.

terms, icons are not. Discursive representations are "inferentially promiscuous" (Stich, 1978), which means that they can figure as premises in logical transitions.

Moreover, lack of logical form renders iconic representations a-rational; they are neither rational nor irrational—the concept of rationality simply does not apply to them. Relations between iconic representations are not logical; these relations are usually understood as a causal chain of associations. An image of a mother can evoke a memory image of a family home, but the link between these two images is not a matter of a logical consequence. We can speak of the temporal or causal sequence of images but rationality is not based on temporal or causal links. It is a matter of following logical rules and reasons. Therefore, iconic representations cannot be rational, and if someone like Frege thinks that the core of thinking is rational thinking, then icons cannot be constituents of thoughts.

Two remarks are required here. First, it may seem that Frege's argument can be easily refuted by pointing out straightforward counterexamples. For instance, if I want to negate that John has red hair, I can depict him as blond. If I depict a green and a red apple, I express an alternative of a green and red apple. If one places two pictures next to each other, much like in a comic book, then one can say that their content is conjoined or implies one another (Westerhoff, 2005).

These examples are, however, misleading. The role of logical form is determining the truth-conditions of its elements. In the case of iconic representations, truth-conditions cannot be determined. Having a picture of John with blond is either a negation of having red hair or black hair, etc., for the content of not-redheaded is not simply being blond but an infinite alternative of a form: being blond or having black hair, or having green hair, etc. No image can represent infinitely many properties.

By the same token, conjunction, disjunction, and implication do not simply represent a sequence of elements; they set up a logical link between them. In the case of two pictures, there is no way to determine the nature of this link—whether it is a temporal sequence, causal link, spatial transformation, or if it is simply a set of two unrelated pictures. In all these cases, the pictorial form is the same.

Let me illustrate these problems with the mental model theory of reasoning (e.g., Barrouillet et al., 2000; Byrne, 2005; Byrne and Johnson-Laird, 1989; Johnson-Laird, 1983). A mental model is a schematic representation of a possible state of affairs. It represents the elements of a set as well as possible spatial and causal relations between the relevant elements of the set. Manipulation of the spatial and causal properties of a model allows one to reason about the properties of the set's element. As an example, let us try to solve the following syllogism:

(1) Some artists are beekeepers.
All beekeepers are chemists.
What follows?

To do this task, we can form a mental model of an artist who is at the same time a beekeeper and a chemist. The task is easy: some artists are chemists.

On the surface, it may be tempting to interpret mental models as iconic, predominantly visual, representations (e.g., Johnson-Laird, 1983). The idea is that we visualise the elements of the syllogism that can help us to solve it. However, visual models are not logical structures (e.g., Hintikka, 1987; Johnson-Laird, 1998; Knauff and Johnson-Laird, 2002). They lack the generality and precision that is required by logical operations. For instance, notice that the same mental model can be used in the cases of the following distinct syllogisms:

(2) Some artists are beekeepers.
All beekeepers are chemists.
Ergo: All artists are chemists.

(3) Some artists are beekeepers.
All beekeepers are chemists.
Ergo: Some artists are chemists.

There is also nothing that would separate this reasoning from fallacious reasoning, such as:

(4) Some artists are beekeepers.
All beekeepers are chemists.
Ergo: All artists are beekeepers.

The problem is that it is impossible to depict the difference between the claim that $\exists x\mathrm{P}(x)$ and $\forall x\mathrm{P}(x)$. Therefore, iconic representations lack logical form.

Second, one might object that iconic representations can exhibit systematicity. For one thing, map-like representations seem to be systematic (e.g., Camp, 2007; Braddon-Mitchell, Jackson, 1996). A part of a map that represents that London is west of Berlin also represents that Berlin is east of London. For another, as Matthen (2005) notes criticizing Evans' Generality Constraint, if one can imagine a blue circle and a red square, then one can imagine a red circle and a blue square. In other words, if a representational system is capable of representing multiple features together, then it can represent different configurations of these features.

These objections, however, miss the mark, for the systematicity of discursive representations comes paired with compositionality. For discursive representations to be systematic, we have to be able to distinguish between the meanings of the constituents and the meanings of the complex structures they form. For instance, the thought that John loves Mary is built out of the concepts John, Mary, and love, which can be distinguished as separate semantical units. In the case of iconic representations, such separation cannot be carried out, for they lack canonical decomposition.

Does the analog-digital distinction help us to understand the iconic-discursive distinction? It seems that it does not, as the iconic-discursive distinction does not overlap with the analog-digital one.

For one thing, as von Neumann (1958) demonstrates, discursive operations, such as computations, can be functions of digital and analog processes. From the hardware point of view, discursive representations can be encoded either in digital or analog format. Thus, the distinction between analog and digital format is irrelevant for determining whether we are dealing with an iconic or discursive representational system. For another, this distinction falsely implies that there are no digital iconic representations. There obviously are. Therefore, being analog or digital is neither necessary nor sufficient for being an iconic or discursive representation, which sometimes leads to the conclusion that the analog-digital distinction is only notational (e.g., Johnson, 2015; Szabo, 2012).

However, the difference between iconic and discursive representational systems can be described in terms of different ways iconic and discursive information is processed in the mind. In the next section, I present an alternative understanding of the concept of representational format.

2. Mental Representation Format as a Way of Processing Information

There are at least two marks that distinguish the structure of iconic and discursive representations. Concerning the first, mechanisms of information processing in iconic representations are domain-specific. The domain-specificity of iconic representations can be understood as a joint alternative of two theses. First, it means that the mechanism of information processing varies depending on the nature of the vehicle of representation. Second, mechanisms of information processing depend on the modality of representation. In contrast, mechanisms of information processing of discursive representations are domain-general, which means here that they do not depend on the features of the vehicle of representation nor the modality of representation.

Concerning the second mark, iconic and discursive representations employ different predicative functions. The structure of iconic content is organised non-hierarchically and is based on holistic data. In contrast, the structure of discursive content is organised hierarchically and is based on discrete chunks of information.

The mechanisms of information processing are domain-specific if the operations defined on the elements of the structure depend on the area of application. The relevant mechanisms are domain-general if the operations do not depend on the area they are applied to. For instance, rules of addition are domain-general; regardless of what one is adding, the rules are the same. In contrast, heuristic rules are domain-specific, for there is no general heuristic that could help us solve every type of cognitive task. Depending on what one is trying to solve, one uses different heuristics.

The mechanisms of information processing in iconic representations are domain-specific. This claim consists of a joint alternative of two theses. For one thing, the mechanism of information processing depends on the features of the vehicle; for another, it depends on the modality of representation. Let us dub them “vehicle-specificity” and the “modality-specificity”, respectively. I distinguish between vehicle-specificity and modality-specificity mostly because I do not want to settle whether discursive representations are amodal here (Prinz, 2002). In other words, one can hold that discursive representations, such as concepts, are modally-specific; thus, modality-specificity is not necessarily a valid criterion for distinguishing between iconic and discursive representations. However, acknowledging vehicle-specificity as a criterion of iconic format does not imply that one has to acknowledge modality-specificity as a relevant criterion (although the implication is the other way round).

Vehicle-specificity means that the features of the vehicle of representation determine the way we process the information. First, access to the information varies depending on whether the information is displayed on an external or internal vehicle of representation. External vehicles of representations are central for using pictures, maps, diagrams, or gestures. Internal vehicles of representations are central for mental imagery and perception. Second, access to the information varies depending on the way information is displayed on the vehicle of representation.

The distinction between internal and external vehicles of representation corresponds to different mechanisms for how iconic and discursive information is processed in the mind. Let us illustrate it with a mental imagery example. In comparison to external images, it is widely believed that the content of mental images is not subject to interpretation but is displayed as already interpreted (e.g., Chambers, Reisberg, 1985; Reisberg, 1996; Reisberg, Heuer, 2005; Sartre, 1962; Slezak, 1995). In contrast, we can always reinterpret the content of picture perception. It means that the meaning of mental images is fixed, while the meaning of external images may change accordingly to the way we perceive the image.

Two clarifications are needed. The inability to reinterpret mental images is subject to scientific dispute. Contrary to the classic positions in cognitive psychology represented, for instance, by Chambers and Reisberg (1985), it seems that we can reinterpret the content of mental images. However, the reinterpretation of mental images is more cognitively loaded and less efficient. For example, we can reinterpret so-called bistable figures (such as the duck-rabbit picture) displayed on an external and an internal representation but with significant differences in the effectiveness of solving the task (Mast, Kosslyn, 2002; Kamermans et al., 2019). That suggests, first, that the difference between mental imagery and picture perception is vaguer than we previously thought. It is not a difference in kind, rather a difference in degree. This fact is easier to understand if we assume that mental imagery and pictorial perception share the same format but that the format is vehicle-specific. Second, it indicates that the descriptivist positions (e.g., Pylyshyn, 1973) that assume only the discursive format of representation in the imagery debate are wrong, for if mental images were encoded in a discursive format, then it would be difficult to explain why mental images are subject to reinterpretation.

Moreover, it is important to keep in mind that even proponents of the pictorial theory of mental imagery, such as Kosslyn, have never claimed that the nature of mental images is the same as picture perception. At most, they speak of the quasi-perceptual nature of imagery. In other words, even if it is not clear what the difference between perception and mental imagery is, no one claims that there is no such difference. For instance, as was pointed out by Hinton (1979),[6] if we form a mental image of nine letters put randomly into a 3×3 grid, it is difficult for us to read the imagined string of letters. However, it is a trivial task if we write down the same letters on paper. Thus, even if information encoded in internal representation is processed similarly to the way information encoded in external representation is processed, they are not of the same processes (Ittelson, 1996). It does not mean that the format is different. It means that the format is vehicle-specific.

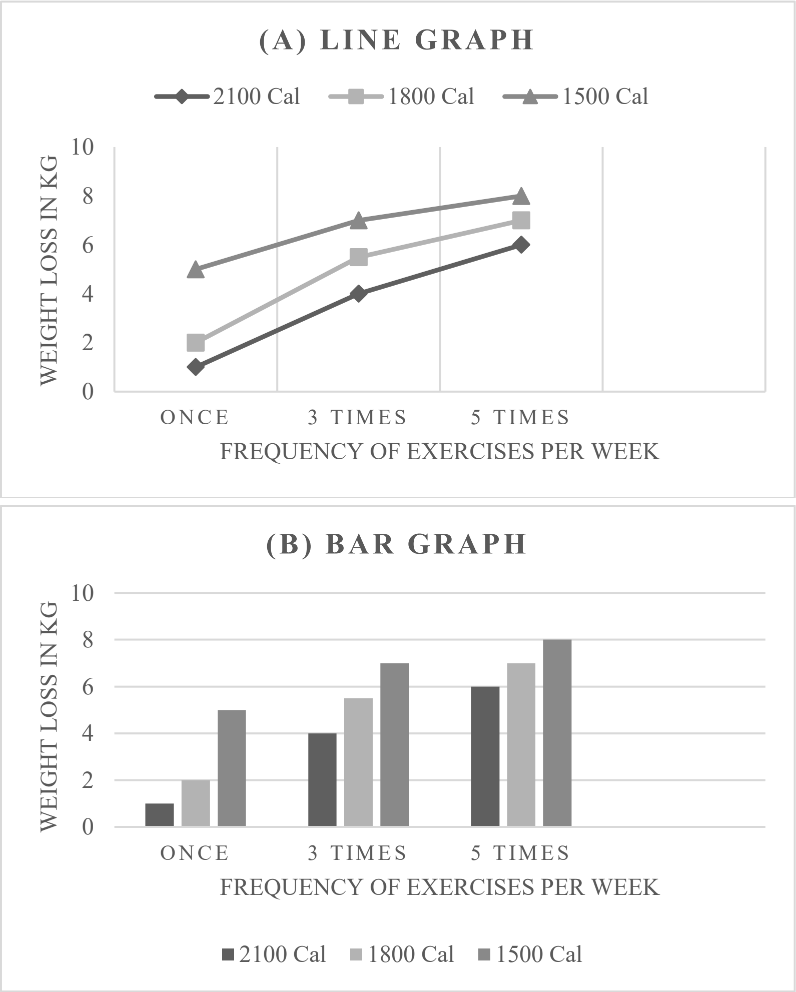
Furthermore, the different ways in which a piece of information is displayed on the vehicle of representation affects the accessibility of the piece of information. For instance, a line graph (a) and a bar graph (b), as shown in Figure 1, can represent the same information but in different ways. Graph (a) makes it easier to understand the relationship between the amount of exercise and weight loss. The data is connected by an increasing function, whereas graph (b) makes it easier to understand the relationship between exercise and the number of calories burned because the bars comparing the data of calories and the amount of exercise are closer to one another.

Broadly speaking, we can assume that the features of the vehicle can affect the mechanisms of information processing in iconic representations. The features of the vehicle do not merely provide input for certain internal processes. The interpretation of the same content displayed on different types of vehicles involves separate mental processes. In other words, the type of vehicle of representation determines what information we have access to and what mental processes are involved in processing that information. Using Larkin and Simon's (1987) formulation, iconic representations are computationally inequivalent due to the type of vehicle of representation.

[6] Kosslyn (1994) argues that Hinton's results indicate the mere fact that mental images are displayed in the mind's eye too short, and, therefore, the results show only how memory limitations affect imagery. However, the time that is not sufficient to solve Hinton's tasks is the same time that suffices to solve mental rotation and mental zooming tasks in Kosslyn's classic research on mental imagery. Therefore, the time for which information is available does not seem to be a relevant factor.

Figure 1

Differences in Information Processing in Graphs



Note. Line graphs (a) and bar charts (b) convey the same information but in a different way, affecting the accessibility of the information.

In contrast, in discursive representations, access to the content of the belief that snow is white is the same as access to the content of the sentence “snow is white”. The mechanism for processing the information that snow is white represented by the thought snow is white and represented by the sentence “snow is white” is the same because the propositions contained in the belief and expressed in the sentence are the same. In other words, the same predicative function predicating white of snow is expressed in both the internal and external representation.

Moreover, there is shared meaning in the case of different external representations. The expression "snow is white" expresses the same information as its equivalent in German "Schnee ist Weiss". The same content can be expressed by two syntactically different representations. The cognitive content of the sentence "snow is white" is the same as "the colour of snow is white".

We face a similar situation in reasoning. For instance, the function 2+2=4 is carried out in the same way regardless of whether it is done in one's head or on paper. The reasoning is the same because the cognitive content is the same. This means that the information is processed in the same way because the operations on the sets are the same. The representations are computationally equivalent. Thus, content is not affected by differences in the features of vehicles of discursive representations.

One remark is required. From the given description, it does not follow that there is no difference in cognitive access between the mental content inside the mind and the content expressed by an external (e.g., linguistic) representation of a thought. A thesis of this type would be false for obvious reasons. For example, calculations performed on a piece of paper may be cognitively less loaded than those performed in thought. However, this does not mean that the structure of information is different. To use a computer metaphor, the .jpg format can be supported by more or less computationally efficient computers. The type of computer, however, does not affect the format of the file.

The second way of understanding the domain-specificity of the iconic format of representation refers to the concept of modality-specificity. The mechanisms of information processing depend on the modality of representation. For instance, a visual and gustatory representation of wine are two different systems of representation—the information is processed differently. In contrast, discursive representations are amodal—discursive representations of the colour and taste of wine are different in content but the information corresponding to the the colour and taste is processed in the same way.

Processing iconic information involves different mechanisms that are responsible for processing information of different sensory types (interoceptive, visual, tactile, auditory, etc.), which is associated with activation of different neurobiological systems. This point can be illustrated with individual differences in visual imagery and spatial cognition tasks (e.g., Hegarty, Waller, 2005; Kozhevnikov, Blazhenkova, Becker, 2010). In short, there are large individual differences in tasks where people are asked to imagine a sunny day and tasks which measure spatial abilities, such as when they are asked to imagine the spatial transformation of mental images. These dimensions are uncorrelated (e.g., Kosslyn et al, 1984) and can negatively affect each other. For example, visualising the content of a problem can lower the effectiveness of reasoning in spatial and abstract problem-solving tasks, which is known as the visual impedance effect (e.g., Knauff, Johnson-Laird, 2002).

The details of different mechanisms of information processing here are of less importance. I only wish to emphasize the fact that iconic representations are modality-specific. In contrast, discursive representations are amodal. The proposition snow is white remains the same when it is expressed in a spoken or written form, in English or in German.

The other detail that distinguishes iconic and discursive representations concerns the problem of predication. It is believed that discursive representations, such as beliefs, have a predicative nature. When I have a belief that snow is white, I attribute the property white to the object snow, where the terms "snow" and "white" work as arguments of a predicative function expressed in the proposition. Thanks to the predicative nature of the proposition, we can distinguish the proposition from a list of terms. The proposition snow is white differs from the list of terms "snow", "white", and "is" because the proposition has a predicative structure that carries the denotations of the terms into a truth-value (e.g., Rescorla, 2009), while the list of terms does not. It means that the proposition snow is white can say something about the world, while the list of terms cannot.

It may seem that iconic representations can be predicative too (e.g., Blumson, 2012; Matthen, 2005). If I form an image of a red triangle, I attribute the property of redness to the triangle. And the image of a red triangle can be distinguished from the conjunction made of an image of a triangle and an image of a red patch. However, to assess whether or not we are dealing with equivocation here, we must have a clear understanding of what predication is.

First, in the case of discursive representations, predication is based on a compositional and combinatorial mechanism. The compositional and combinatorial character of a discursive representational system means that if I possess the propositions snow is white and a triangle is red, I can form the structurally similar propositions snow is red and a triangle is white. Moreover, if I know that snow is white is true, I can infer that the proposition snow has a colour is also true. It means that an output of one operation of predication can be an input of another higher-order operation. These operations are hierarchically organised (e.g., Camp, 2018). From the proposition snow is white, I can infer that snow has a colour, but from the proposition snow has a colour I cannot infer that snow is white.

Second, the output and input information comes in discrete chunks. It means that the proposition snow is white attributes the property whiteness and no other property to snow. It says nothing about the hue of the colour or the shape of snow, for there is one-to-one correspondence between a vehicle and a content. Every chunk of information needs a separate vehicle to be expressed. Thus, discursive representations are hierarchically organised and are based on a structure made of discrete chunks of information.

In contrast, iconic representation is neither hierarchically organised, nor is it based on discrete chunks of information. Hierarchical organisation of representational structure means, for instance, that the proposition snow is white implies snow has a colour but not another way round. In the case of iconic representation, both pieces of information are processed simultaneously. An image of white snow represents both that snow is white and that snow has a colour. That is why it does not matter from which point we start analysing an image of white snow—whether from thinking of snow as having a colour or of snow being white—both starting points lead to the same result.

The non-hierarchical structure of iconic representations can be illustrated with the help of maps. On maps, all of the pieces of information about the locations of objects are displayed simultaneously. The information that London is west of Berlin is simultaneously displayed on the map with the information that London is west of Warsaw. In the case of discursive representation, the information that London is west of Berlin does not contain the information that London is west of Warsaw. It must be inferred from the conjunction of the propositions Berlin is west of Warsaw and London is west of Belin. In contrast, a map displays information about all of the possible spatial relations simultaneously.

The non-hierarchical organisation of information processing is often confused with the holistic nature of the components of iconic structure. These two concepts, however, have to be separated, since we can have non-hierarchically organised processes based on non-holistic components. For instance, parallel computing is non-hierarchically organised and is based on discrete chunks of information. However, there is no clear account of what "holistic representation" means. There are at least three interpretations of this term.

First, the concept of the holistic nature of iconic representation is often interpreted as indicating the fact that iconic representations are informationally rich (Dretske, 1981; Kitcher, Varzi, 2000) and fine-grained (Tye, 2005). It means, first, that iconic content conveys so much information that it cannot plausibly be expressed with a finite set of propositions and, second, that iconic content is detailed and determined. In contrast, the content of discursive representations is general and abstract. For instance, seeing white snow is having an experience of a determined shade of white—thinking that snow is white does not determine the shade of the colour.

Although the concepts of information richness and being fine-grained seem intuitive, they are far from clear. First, there are discursive representations that are rich in content, such as the symbol π, and iconic representations that are informatively primitive, such as an image of a dot. Moreover, even if the richness of detail we are dealing with cannot plausibly be expressed by a finite set of propositions, it does not mean that it is impossible. A potentially infinite set of complex propositions can express any amount of information. Second, discursive content can be more fine-grained than iconic content. Pictures of aqua and cyan objects are often not detailed enough to see the difference between them; the propositions x is aqua and y is cyan are. Therefore, informative richness and fineness of grain do not determine whether we are dealing with iconic or discursive representations.

Second, we can interpret the concept of holistic representation as indicating the fact that pieces of information are entangled in a representation. For instance, Camp (2018) understands the holistic nature of information processing in maps as a matter of structural linkage of the pieces of information. It means that changing the informational content of one of the map's elements changes the informational content of every other element. For example, moving the position of London on a map changes its distance to every other point on the map. In contrast, changing the informational content of *p* does not have to change anything about the content of *q*.

However, the problem is the scope of this thesis. Camp's argument certainly applies to information concerning spatial relations. However, it only says that spatial properties are relational, which is trivial. It does not apply to non-spatial information. For instance, changing the size of a circle representing London's population does not change the information about the size of a circle representing the population of Berlin.

Third, the holistic nature of representation can be interpreted (as it is interpreted here) as indicating the relation between the content and the vehicle of representation. To my knowledge, the first person to draw attention to this idea was Kazimierz Twardowski (1965), who ascribed the feature of concreteness to imaginings. He understood concreteness as the combination of multiple properties in a single representation. Similarly, the concept of holism is sometimes understood (e.g., Green, Quilty-Dunn, 2017; Kulvicki, 2020) as the thesis that multiple pieces of information expressing the content of a representation are assigned to the same vehicle of representation. It means that there is no separate vehicle for every chunk of information corresponding to different representational properties. In other words, there is no one-to-one correspondence between parts of the information and parts of the vehicle. For instance, the part of an icon that represents the colour of a triangle is the same part that represents its shape and location. Likewise, in the case of maps. The part of a topographic map that represents the location of London on the map represents its height above sea level too, along with a host of other things.

Thus, although iconic and discursive representations are described as expressing predicate functions, the way in which they process information is different. They express distinct predicative functions.

To sum up, iconic representations can be distinguished from discursive representations based on differences in the structure of information processing. The structure of iconic representations (or their format) is domain-specific; it processes information in a non-hierarchical manner and is based on holistic components. The structure of discursive representations is domain-general; it processes information hierarchically and is based on discrete chunks of information.

3. Canonical Decomposition and Lack of Logical Form

In the first section, I claimed that the distinction between different formats of representation should be able to provide at least a partial explanation of why iconic representations lack canonical decomposition and logical form, and why discursive representations are canonically decomposable and inferentially promiscuous. In this section, I explain how the description of the representational structure I presented in the second section can help us to better address these questions.

First, Fodor's argument regarding the lack of canonical decomposition holds that iconic content cannot be decomposed into canonically distinguished parts. However, in this minimal form, the argument is clearly false, for the claim that iconic representations lack syntactic structure works as both a premise and the conclusion of the argument.

To fully present the structure of Fodor's argument, it is necessary to supplement it with a metaphysical premise concerning the format of iconic representation. According to the metaphysical premise, iconic representations have no structure linking the represented properties to the vehicle of representation, and therefore they cannot be canonically decomposed.

Let me illustrate the metaphysical premise with an image of a red square. The image of a red square can represent the content of the concept red or square, as well as the content of the proposition some squares are red. Yet, if we have a mental image of a red square, there is no way to determine whether we are thinking of the concept square or the content of the proposition some squares are red. Discursive representations can distinguish between concepts and structures composed of concepts, such as propositions. Iconic representations are unable to do so.

Likewise, in the case of maps. A map representing that Warsaw is east of Berlin represents Warsaw, Berlin and the fact that Warsaw is east of Berlin. We can distinguish between the representations Warsaw and Berlin and the representation Warsaw is east of Berlin only if we have the concept east. The concept east allows us to isolate the spatial relation property from all other represented properties. However, this means that we need a representation that can isolate a particular bit of information and assign it to the relevant representation vehicle. The discursive representation east does exactly that.

Iconic representations are unable to do so because there is no one-to-one correspondence between the information and its vehicle. As I claimed, the structure of iconic representations is based on holistic components. Multiple pieces of information are displayed on the same vehicle of representation. An image of a red square represents both redness and squareness, as well as the fact that some squares are red. In other words, iconic representations lack constituent structures and cannot be canonically decomposed, for there is no way to assign a single bit of information to a corresponding distinct vehicle of information.

Moreover, the domain-specificity of iconic representations renders them unable to specify what the canonical decomposition of iconic representations could be. Compare spatial and non-spatial iconic representations. Although we can cut a map into spatial pieces, not all iconic representations have spatial parts. Gustatory representations do not. For instance, the taste of a meal can be described as savory or sweet, but it is not dividable into spatial pieces. Therefore, there is no general way to decompose iconic representations.

In contrast, discursive representations can be decomposed canonically since their structure is domain-general. The proposition snow is white can be divided into canonically isolated parts regardless of whether it is expressed in English or in German.

Second, the systematicity of discursive representations means that they are inferentially promiscuous (Peacocke, 1992; Stich, 1978). For instance, beliefs and thoughts can figure as premises in inferences. Discursive representations have a logical form and can be modelled according to the rules of logic. Yet, according to Frege's argument, iconic representations lack logical form. They cannot be inferred or negated. They are not inferentially promiscuous. Why is that so?

Inferential promiscuity requires propositional structure. It involves a kind of relation between the vehicle of representation and the content. To infer from *a* is *F* and if *a* is *F*, then *a* is *G* that *a* is *G*, one has to be able to assign distinct contents to the vehicles of representation expressed by logical variables. Moreover, the chunks of information have to be hierarchically organised. From the thought that *a* is *G* and that if *a* is *F*, then *a* is *G* I cannot infer that *a* is *F*. From the thought that snow is white I can infer that it has a colour, but from snow has a colour I cannot infer that it is white. Discursive representations are hierarchically organised.

In contrast, images are organised non-hierarchically. The information is processed simultaneously. When I see a picture of white snow, I see that it has a colour; when I see a colourful picture, I can see the specific colour of the picture.

Moreover, inferential promiscuity requires a representational structure that can abstract from the nature of the vehicle of representation. For instance, the reasoning if *A* then *B* and *A* then *B* is correct regardless of whether the reasoning is conducted in the mind or on paper. Discursive representations meet this requirement since they are domain-general. In contrast, iconic representations are domain-specific. The nature of the vehicle of iconic representation affects the way information is processed. For instance, tasting wine and imagining tasting wine are informationally two different representations.

To sum up, iconic representations lack syntactic structure and do not meet the requirements of the Generality Constraint. They are neither systematic nor canonically decomposable. These facts are easier to understand if we hold that iconic representations are domain-specific, that they process information in a non-hierarchical fashion, and that their structure is based on holistic components. In contrast, discursive representations are domain-general, they process information hierarchically, and their structure is based on discrete elements. Thus, discursive representations are systematic and canonically decomposable.

The argument presented here does not show that the distinction between analog and digital representational systems is useless. However, it demonstrates that this distinction is insufficient for distinguishing perception and thought, for it does not provide any explanation of the source of the differences between iconic and discursive representations. In contrast, thinking of iconic and discursive representations in terms of the way they structure information helps us to better understand why they differ. According to the view presented here, the different functional properties of iconic and discursive representations follow from different informational structures.

REFERENCES

Barrouillet, P., Grosset, N., Lecas, J.-F. (2000). Conditional Reasoning by Mental Models: Chronometric and Developmental Evidence. *Cognition*, *75*(3), 237–266.

Beck, J. (2015). Analogue Magnitude Representations: A Philosophical Introduction. *The British Journal for the Philosophy of Science*, *66*(4), 829–855.

Blumson, B. (2012). Mental Maps. *Philosophy and Phenomenological Research*, *85*(2), 413–434.

Braddon-Mitchell, D., Jackson, F. (1996). *Philosophy of Mind and Cognition*. Oxford: Blackwell.

Byrne, R. M. J. (2005). *The Rational Imagination: How People Create Counterfactual Alternatives to Reality*. Cambridge, MA: MIT Press.

Byrne, R. M. J., Johnson-Laird, P. N. (1989). Spatial Reasoning. *Journal of Memory and Language*, *28*, 564–575.

Camp, E. (2004). The Generality Constraint and Categorial Restrictions. *Philosophical Quarterly*, *54*(215), 209–231.

Camp, E. (2007). Thinking with Maps. In J. Hawthorne (Ed.), *Philosophical Perspectives 21: Philosophy of Mind* (pp. 145–182). Oxford: Wiley-Blackwell.

Camp, E. (2018). Why Maps Are Not Propositional. In A. Grzankowski, M. Montague (Eds.), *Non-Propositional Intentionality* (pp. 19–45). Oxford: Oxford University Press.

Casati, R., Giardino, V. (2013). Public Representation and Indeterminacies of Perspectival Content. In Z. Kondor (Ed.), *Enacting Images* (pp. 111–126). Köln: Herbert von Halem Verlag.

Chambers, D., Reisberg, D. (1985). Can Mental Images Be Ambiguous? *Journal of Experimental Psychology: Human Perception and Performance*, *11*(3), 317–328.

Crane, T. (2009). Is Perception a Propositional Attitude? *The Philosophical Quarterly*, *59*, 452–469.

Devitt, M. (2006). *Ignorance of Language*. Oxford: Oxford University Press.

Dretske, F. (1969). *Seeing and Knowing*. Chicago: University of Chicago Press.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press.

Fodor, J. (2007). The Revenge of the Given. In B. P. McLaughlin, J. D. Cohen (Eds.), *Contemporary Debates in Philosophy of Mind* (pp. 105–116). Oxford: Basil Blackwell.

Fodor, J. (2008). *LOT 2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Fodor, J., Pylyshyn, Z. (1981). How Direct Is Visual Perception? Some Reflections on Gibson's 'Ecological Approach'. *Cognition*, *9*, 207–246.

Fodor, J., Pylyshyn, Z. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition*, *28*(1–2), 3–71.

Frege, G. (1984). Thoughts. In B. McGuinness (Ed.), *Collected Papers on Mathematics, Logic, and Philosophy* (pp. 351–372). Oxford: Basil Blackwell.

Goodman, N. (1976). *Languages of Art* (2nd Ed.). Indianapolis: Hackett.

Green, E. J., Quilty-Dunn, J. (2017). What Is an Object File? *The British Journal for the Philosophy of Science*. doi:10.1093/bjps/axx055

Haugeland, J. (1998). *Having Thought: Essays in the Metaphysics of Mind*. Cambridge: Harvard University Press.

Heck, R. G. (2000). Nonconceptual Content and the 'Space of Reasons'. *The Philosophical Review*, *109*, 483–523.

Heck, R. G. (2007). Are There Different Kinds of Content? In B. P. McLaughlin, J. Cohen (Eds.), *Contemporary Debates in Philosophy of Mind* (pp. 117–138). Oxford: Blackwell.

Hegarty, M., Waller, D. A. (2005). Individual Differences in Spatial Abilities. In P. Shah, A. Miyake (Eds.), *The Cambridge Handbook of Visuospatial Thinking* (pp. 121–169). New York: Cambridge University Press.

Hintikka, J. (1987). Mental Models, Semantical Games, and Varieties of Intelligence. In L. Vaina (Ed.), *Matters of Intelligence: Conceptual Structures in Cognitive Neuroscience* (pp. 197–215). Dordrecht: D. Reidel.

Hinton, G. (1979). Some Demonstrations of the Effects of Structural Descriptions in Mental Imagery. *Cognitive Science*, *3*(3), 231–250.

Ittelson, W. H. (1996). Visual Perception of Markings. *Psychonomic Bulletin & Review*, *3*(2), 171–187.

Johnson, K. (2015). Maps, Languages, and Manguages: Rival Cognitive Architectures? *Philosophical Psychology*, *28*(6), 815–836.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge: Harvard University Press.

Johnson-Laird, P. N. (1998). Imagery, Visualization, and Thinking. In J. Hochberg (Ed.), *Perception and Cognition at the Century's End* (pp. 441–467). San Diego: Academic Press.

Kamermans, K. L., Pouw, W., Mast, F. W., Paas, F. (2019). Reinterpretation in Visual Imagery Is Possible Without Visual Cues: A Validation of Previous Research. *Psychological Research: An International Journal of Perception, Attention, Memory and Action*, *83*(6), 1237–1250.

Kitcher, P., Varzi, A. (2000). Some Pictures are Worth 2[aleph]0 Sentences. *Philosophy*, *75*(3), 377–381.

Kosslyn, S. M. (1980). *Image and Mind*. Cambridge, MA: Harvard University Press.

Kosslyn, S. M. (1994). *Image and Brain: The Resolution of the Imagery Debate*. Cambridge, MA: MIT Press.

Kosslyn, S. M., Brunn, J., Cave, K. R., Wallach, R. W. (1984). Individual Differences in Mental Imagery Ability: A Computational Analysis. *Cognition*, *18*, 195–243.

Kozhevnikov, M., Blazhenkova, O., Becker, M. (2010). Trade-off in Object Versus Spatial Visualization Abilities: Restriction in the Development of Visual-Processing Resources. *Psychonomic Bulletin & Review*, *17*(1), 29–35.

Knauff, M., Johnson-Laird, P. N. (2002). Visual Imagery Can Impede Reasoning. *Memory and Cognition*, *30*, 363–371.

Kulvicki, J. (2006). *On Images: Their Structure and Content*. Oxford: Clarendon Press.

Kulvicki, J. (2020). *Modelling the Meanings of Pictures: Depiction and the Philosophy of Language*. Oxford: Oxford University Press.

Larkin, J. H., Simon, H. A. (1987). Why a Diagram Is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, *11*, 65–99.

Lewis, D. (1971). Analog and Digital. *Noûs*, *5*, 321–328.

Lorenzo, G., Rubiera, E. (2019). On Iconic-Discursive Representations: Do They Bring Us Closer to a Humean Representational Mind? *Biosemiotics*, *12*, 423–439.

Maley, C. J. (2011). Analog and Digital, Continuous and Discrete. *Philosophical Studies*, *155*, 117–131.

Mast, F. W., Kosslyn, S. M. (2002). Visual Mental Images Can Be Ambiguous: Insights From Individual Differences in Spatial Transformation Abilities. *Cognition*, *86*, 57–70.

Matthen, M. (2005). *Seeing, Doing, and Knowing: A Philosophical Theory of Sense Perception*. Oxford: Clarendon Press.

McDowell, J. (1996). *Mind and World*. Cambridge, MA: Harvard University Press.

McGinn, C. (1989). *Mental Content*. Oxford: Blackwell Publishers.

Pagin, P., Westerståhl, D. (2010). Compositionality I: Definitions and Variants. *Philosophy Compass*, *5*, 250–264.

Peacocke, C. (1989). Perceptual Content. In J. Almog, J. Perry, H. Wettstein (Eds.), *Themes from Kaplan* (pp. 297–329). New York: Oxford University Press.

Peacocke, C. (1992). *A Study of Concepts*. Cambridge, MA: MIT Press.

Prinz, J. J. (2002). *Furnishing the Mind. Concepts and their Perceptual Basis*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (1973). What the Mind's Eye Tells the Mind's Brain: A Critique of Mental Imagery. *Psychological Bulletin*, *80*, 1–25.

Pylyshyn, Z. W. (1981). The Imagery Debate: Analogue Media Versus Tacit Knowledge. *Psychological Review*, *88*(1), 16–45.

Quilty-Dunn, J. (2016). Iconicity and the Format of Perception. *Journal of Consciousness Studies*, *23*(3–4), 255–263.

Quilty-Dunn, J. (2020). Perceptual Pluralism. *Noûs*, *54*(4), 807–838.

Reisberg, D. (1996). The Nonambiguity of Mental Images. In C. Cornoldi, R. H. Logie, M. A. Brandimonte, G. Kaufmann, D. Reisberg (Eds.), *Stretching the Imagination: Representation and Transformation in Mental Imagery* (pp. 119–172). New York: Oxford University Press.

Reisberg, D., Heuer, F. (2005). Visuospatial Images. In P. Shah, A. Miyake (Eds.), *The Cambridge Handbook of Visuospatial Thinking* (pp. 35–80). New York: Cambridge University Press.

Rescorla, M. (2009). Predication and Cartographic Representation. *Synthese*, *169*(1), 175–200.

Sainsbury, R. M. (2005). *Reference without Referents*. Oxford: OUP.

Sartre, J.-P. (1962). *Imagination: A Psychological Critique*. Ann Arbor: University of Michigan Press.

Sellars, W. (1997). *Empiricism and the Philosophy of Mind*. Harvard: Harvard University Press.

Slezak, P. (1995). The 'Philosophical' Case Against Visual Imagery. In P. Slezak, T. Caelli, R. Clark (Eds.), *Perspectives on Cognitive Science: Theories, Experiments and Foundations* (pp. 237–271). Norwood: Ablex Publishing.

Sober, E. (1976). Mental Representations. *Synthese*, *33*, 101–148.

Stich, S. P. (1978). Beliefs and Subdoxastic States. *Philosophy of Science*, *45*(4), 499–518.

Szabo, Z. (2012). The Case for Compositionality. In M. Werning, W. Hinzen, E. Machery (Eds.), *The Oxford Handbook of Compositionality* (pp. 64–80). Oxford: Oxford University Press.

Travis, C. (2013). *Perception—Essays After Frege*. Oxford: Oxford University Press.

Twardowski, K. (1965). Wyobrażenia i pojęcia. In K. Twardowski, *Wybrane Pisma Filozoficzne* (pp. 114–197). Warsaw: PWN.

Tye, M. (2005). Non-Conceptual Content, Richness, and Fineness of Grain. In T. Gendler, J. Hawthorne (Eds.), *Perceptual Experience* (pp. 504–526). Oxford: Oxford University Press.

Von Neumann, J. (1958). *The Computer and the Brain*. New Haven: Yale University Press.

Westerhoff, J. (2005). Logical Relations between Pictures. *Journal of Philosophy*, *102*(12), 603–623.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXV, nr 1 (2021), s. 95–106
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxv1.06
Article

MACIEJ GŁOWACKI [*]

ON PRODUCTION AND USE OF TOKENS OF "I"[1]

SUMMARY: In this paper, I analyze the semantics of the first person pronoun "I" from the perspective of the user/producer distinction. In the first part of the paper, I describe the Simple View (SV) and propose three interpretations of its thesis (following de Gaynesford, 2006). In the second part, I analyze the notions of use and production of a linguistic token. In the next part, I show that all of the interpretations of SV are sensitive to counterexamples. In the end, I discuss possible answers of the proponents of SV and argue against them. The first aim of this paper is to show that SV is wrong, and the second is to convince the reader that the user/producer distinction is of high importance in the philosophy of language.

KEYWORDS: indexical expression, pure indexicals, I, user, producer, use, production, de Gaynesford.

1. Introduction

What kind of facts constitute the reference of "I"? When I utter "I am hungry", it refers to me. When you write "I don't like the government" on the wall, it refers to you. But in virtue of what is it so? The orthodox answer to this question points toward the simple facts about the context: my utterance of "I" refers to me because I am the agent of the described context; your inscription on the wall refers to you since you are the agent of this context. According to this line of

[*] University of Warsaw, Faculty of Philosophy. E-mail: m.glowacki6@uw.edu.pl. ORCID: 0000-0001-6678-7107.

[1] I want to express my gratitude for their helpful comments on previous versions of the paper to Tadeusz Ciecierski and an anonymous referee of the journal. I also want to thank Zofia Sław and Agnieszka Głowacka for making this text readable.

reasoning, which I will call the Simple View (SV), such simple facts about the context suffice to determine the referent of "I".

In this paper, I argue that the Simple View is ambiguous and, moreover, that it is wrong under each of its readings. In the first part of the paper, I describe the Simple View and propose three plausible interpretations of its thesis. The first interpretation says that the agent of the context is the user of "I", the second states that it is the producer of "I", and the third is that it is either the user or the producer. In the second part, I analyze the notions of use and production of a linguistic token. In the next part, I show that all interpretations of SV are sensitive to counterexamples. In the end, I discuss a possible answer of the proponents of SV and argue against it. The paper has two aims: the first is to show that SV is wrong, and the second is to convince the reader that the user/producer distinction is of high importance in the philosophy of language.

My investigations are inspired and influenced by the book *I: The Meaning of the First Person Term* by Maximillian de Gaynesford (2006).[2] In the book, he argues against the interpretation of "I" as a pure indexical expression and proposes a theory of the semantics of "I" as a prototype of demonstrative.[3] In my paper, I begin with considering his claim about the ambiguity of SV and elaborate on it. I propose an analysis of the notions of production and use of a linguistic token that are important in de Gaynesford's argumentation against the orthodox semantics of "I". I also generalize his counterexamples to SV to adjust them to its different formulations.

Throughout this paper, I assume that the uses of expressions are the bearers of semantic properties, such as reference. By uses I mean acts of using a token of an expression to communicate or to contribute a constituent to the propositional or communicated content.[4] The two alternative bearers of reference—types or tokens of expressions—do not seem plausible candidates. Theories treating types of expressions as having the reference cannot accommodate the variability of reference of indexicals or other context-sensitive expressions. Tokens of expressions have a similar problem: it seems they can change meaning with the change of the occasion of use. I am not saying that there are no unused (in the sense indicated above) tokens that have a reference. I believe though, that they can have a reference in virtue of their future, past, or possible uses. Nevertheless, there is no more I can say here on this matter and I will not argue for it in such a short paper.

[2] The classic monograph on the subject is (Brinck, 1997).

[3] For a discussion of de Gaynesford's claim and arguments, see (Penco, 2021).

[4] Note that this understanding of "use" differs from the Strawsonian one (cf. Strawson, 1950) and semiotic notions of use and usage (Pelc, 1971; Ciecierski, 2021). It seems, however, that we can treat the semiotic notion of use as an equivalence class of uses (in the sense assumed in the present paper) of the equivalence relation of having the same referent.

2. The Simple View

The Simple View says that the simple facts about the context are sufficient to determine the referent of the given use of “I”. By simple facts about the context, we mean answers to the following questions: “What is the time of use?”, “Where is it used?”, “Who is the agent?”.[5] Moreover, proponents of the Simple View usually claim that it is an instantiation of a more general rule: such simple facts about the context constitute the reference of all the so-called pure indexicals.

David Kaplan, in his seminal papers *On the Logic of Demonstratives* (1979) and *Demonstratives* (1989a), provides us with the formal semantic theory of indexicals: both *true demonstratives*—indexicals that need accompanying demonstration to refer (like “this”, “that”, “he”, “over there”), and pure indexicals—indexical expressions that refer solely in virtue of the linguistic meaning of the expression (like “here”, “now” or “I”). Kaplan offers a unified framework for treating both kinds of indexicals. He proposes a distinction between two kinds of meaning of an indexical expression: linguistic meaning (character in Kaplan’s terminology) and content.[6] The character of a demonstrative, like “this”, may be given by “the object indicated by the speaker”. In order to determine the content of “this” in a given situation of its use, we need to have an additional piece of information about the context, namely, information about which object was indicated by the speaker of the context. In the case of pure indexicals, says the Simple View, we do not need more information than the information about the time, the place, or the agent. The character of “I” picks exactly the agent of the utterance. To determine the content of “now”, we need only the time of the utterance, and to determine the content of “here”, we need only the place of the utterance. There is nothing more in the linguistic meaning of these expressions.

The semantic thesis of SV is that the reference of the use of a token of “I” is not only described, but fully determined by the following Simple Rule:[7]

(SR) The referent of the use of “I” is always the agent of the context of the use.

The rule is indeed very simple. The agent is one of the elements of the context and the information about the agent’s identity is the only one we need to determine the reference of any use of “I”. It is quite unusual for a referring expression. Consider e.g., expressions like “this” or “he”. To fix the reference of utterances of these expressions, we need additional elements of the context: at least a demonstration or information about contextual salience. In the case of “I”, or any other pure indexical, we only need one specific element of the context.

[5] Sometimes also a little peculiar question “What world it is?”. In general, we may regard simple facts about the context, as facts concerning chosen subset of contextual coordinates in David Lewis’ terminology from (Lewis, 1970).

[6] To be metaphysically precise: linguistic meaning is a property of expression types, whereas content is a property of concrete uses of expression tokens.

[7] I take the notion of the Simple Rule from (de Gaynesford, 2006).

The Simple View looks very simple but it is ambiguous, as was noted by de Gaynesford (2006, Chapter 2). For, to understand it fully, we have to answer two nontrivial questions:

1. What is the context of use?
2. What is the agent of the context?

The first question arises when one is considering examples of so-called remote utterances (Sidelle, 1991; Briciu, 2017). The second one is due to the possibility of using the same token to express different propositions (Corazza et al., 2002).

There are two sensible answers to the first question: the context of making an actual utterance, inscription, sign, etc. or the intended context of interpretation. This issue becomes vividly noticeable in cases of remote utterances, such as the widely discussed Answering Machine Paradox (Sidelle, 1991).[8] The paradox concerns the semantic evaluation of the statement "I am not here now" recorded on the machine and played during the absence of the person who had recorded it. Such a statement would be true on the second interpretation of context since a person intended the utterance to be interpreted while he or she is not in the place of utterance, but it would be false on the first reading of "context" (Predelli, 1998). I do not want to go deeper into this subject here, but the distinction between contexts of the actual utterance and context of the intended interpretation is worth bearing in mind, while considering the second distinction.

When it comes to the second question, there are also at least two possible answers to it. The agent of the context may be understood as a person who produced the token of an expression (e.g., made a scribble on the wall or produced a sequence of sound waves) or as a person who used the token (e.g., made a political confession using a token of a sentence).[9] In most cases, the distinction has no importance for fixing the referent of "I", but, as we shall see, there are cases in which the distinction is crucial.

Both of the interpretations of the phrase "the agent of the context" can be found in philosophical literature. It is quite commonplace to see statements like: "Whenever [the expression 'I'] is used by a speaker of English, it stands for, or designates, that person" (Barwise, Perry, 1981, p. 670); "'I' refers to the speaker or writer […] of the relevant occurrence of the word 'I'" (Kaplan, 1989, p. 505); "Any token of 'I' refers to whoever produced it" (Campbell, 1994, p. 102).[10]

[8] To the best of my knowledge, a similar problem was first described by Vision (1985).

[9] The phrase "use of the token" is ambiguous. I mean here using it in a "proper way", for communication purposes. We may use a token of a word just to check the microphone before the talk, but it is not the use we are talking about here. Just as we do not count supporting the piano with *The Oxford Handbook of Philosophy of Language* as using it properly.

[10] For a list of similar formulations and a discussion of differences between them see the work of de Gaynesford (2006, p. 36–39)

Often all kinds of such rules are present in the writings of a single philosopher. It seems that it is due rather to the ambiguity of the rule that the authors want to express, than to any serious disagreement between them. For example, "the speaker or writer" from Kaplan's quote is ambiguous between the two readings: the user and the producer.

We might therefore indicate three different interpretations of the rule (SR):

(SR-1) The referent of the use of "I" is always its user.

(SR-2) The referent of the use of "I" is always its producer.

(SR-3) The referent of the use of "I" is always either its user or its producer.[11]

Which is the correct rule describing the reference of the use of "I"? I will argue that neither of them is right. But before that, we shall elucidate the distinction between the user and the producer of the token of an expression.

3. The User/Producer Distinction

As we have seen, there is certain sloppiness in the use of terms "producer" and "user" and the usage of "product" and "use" in philosophical literature. However, there is a reason for this terminological mess in philosophical discussions. The reason is simple: in most cases, it does not matter whether we talk about the user or about the producer of the linguistic token since both notions coincide in the standard examples. When I utter "I am hungry" in a casual conversation, I am both the user and the producer of the given token of the word "I". This use of "I" refers to me and there is no need to ask, whether it is so in virtue of me being a producer of the token of the word "I" or in virtue of me being its user.

It is, however, worth noticing that, although these two notions often overlap, they may also split in some cases. When a painter paints "Phone Oxford 1212 if you wish to complain about me" on hundreds of cars, he produces tokens of that sentence type. A driver who drives one of these cars uses the token to communicate certain content. The painter is the producer of the token which is used by the driver (de Gaynesford, 2006, p. 49). Similarly, if I find a piece of paper on a street with a written token of the sentence "Can you spare a quarter?", I can use it to ask people whether they can spare a quarter, without producing the token of the sentence type (Cappelen, 2011, p. 95). In this example, I am the user of the token and the producer is the person who wrote the sentence on the piece of paper I found.

[11] The third interpretation was not indicated in de Gaynesford's book, but it seems to be a viable option. Moreover, it is not falsified by the counterexamples given in the book. I formulate this option as an exclusive disjunction. It cannot be formulated as a simple disjunction since, as we shall see, in some cases the user and the producer of the token are not the same people. The first person pronoun "I" is a singular referring term and hence has to pick out exactly one element of the context as a reference of the token of "I".

It is sometimes the case that the producer and the user are the same person but the production and the use are made in a different time or place. Consider the case of remote utterances. When on Sunday I write a post-it note for Monday stating that "I am not here today", to inform people that on Monday I will be absent from my room, I produce the token of "I am not here today", but I do not use it. I use it by putting the note on the door of my room. If I used it in the time of producing, I would be lying. But, obviously, no one can call me a liar in virtue of merely producing the token. Similarly, when I record a message for an answering machine: "Hi! I cannot talk right now", I produce the token by recording it, but it is used only when it is listened to by someone who is calling.

Both production and use are intentional actions. The producer and the user are agents performing respective acts. The user is an agent who uses the token to communicate or to contribute a constituent to the propositional or communicated content. The producer is an agent who produces the token *qua* physical object. In normal cases, like in the case of my utterance "I am hungry" during a conversation, an event of my utterance can be described in both ways. I intentionally produce certain sounds using my vocal cords and, at the same time, I use these sounds to express the proposition that I am hungry to my interlocutor.

We may note some obvious differences between the acts of production and use of linguistic tokens. For each token of a linguistic entity, there is only one act of production of the token, but there may be several different uses. For example, the "Do not disturb" sign may be produced once and then be used on several occasions by guests of a hotel. On the other hand, a token may be produced by more than one person in a more complex process of production. Consider for example the giant token of the name "Hollywood", which was probably produced by several different workers.

It is also worth noticing that acts of using and producing a linguistic token have different criteria of success. I can produce a token of a Chinese word 單詞 without any idea what it means, how it is used in a Chinese-speaking community, or even without being able to recognize the different parts that it is composed of. Without knowledge about the usage of this word, it is hard to imagine how I can intentionally use it as a word.

This is not to say that there are no requirements for being a producer of a linguistic token. Such requirements are the matter of a heated debate between philosophers. Some authors claim that the intention to produce the very word 單詞 will be necessary for it being a token of this word (Kaplan, 1990; 2011). Others point toward the requirements of orthographic shape (Katz, 2000; Wetzel, 2009) or the adjustment to the conventions of a given linguistic community (Cappelen, 1999; Hawthorne, Lepore, 2011). I do not want to get deeper into this rabbit hole here. It is, however, worth having in mind that there are requirements for effective production of a linguistic token and that they are distinct from requirements for successful use.

The crucial difference is that users and producers have different aims in their actions. The producer of the linguistic token aims at producing a physical object:

a scribble, a sequence of sound waves, or several Morse signals. He or she produces the token *qua* physical entity. The user, on the other hand, treats the token as possessing semantic properties like meaning, reference, and so on. He or she uses the token as a part of a particular language to express his or her thoughts. Moreover, the user is not aware of the properties of the token *qua* physical object, while using it.[12]

I believe that the precise definition of the use and production of linguistic tokens is a matter of high importance for semantics and metaphysics of language. It is, however, a difficult task that would require deliberations extending the scope of this paper. For our purposes, it suffices to be aware of the distinction. It will help us to show that the Simple View is wrong under each of the interpretations stipulated above.

4. Counterexamples to the Simple View

Proponents of the Simple View state that the reference of the given use of "I" is determined by the simple facts concerning who the agent of the context is. The reference is given by the Simple Rule:

(SR) The referent of the use of the token of "I" is the agent of the context of use.

As we have noticed, there is an ambiguity between the three interpretations of (SR). The agent of the context may be understood as the person who used the token (SR-1) or as the person who produced it (SR-2), or as either of them (SR-3). Now I am going to show counterexamples to all three interpretations.

From our remarks about the notion of a producer of a linguistic token, it should be quite clear that (SR-2) is not a suitable interpretation of (SR). There are many possible cases in which a person is not the referent of the token of "I", despite being its producer. For example, I may speak in court on behalf of a mute person. When I say "I admit to committing this crime" for her, the referent of "I" is she, not me, although I produced the token using my vocal cords (de Gaynesford, 2006, p. 41). In a more extreme scenario, I may be forced to speak by a demon who possessed my poor soul, by a mad scientist, or by a professional hypnotist. In these situations, when I say, "I am capable of possessing his poor soul", the token of "I" that I produced surely does not refer to me. It refers to the possessor, who is using me to express the proposition that he is capable of possessing my poor soul. There are of course more mundane cases in which the producer is not a referent of the token of "I". Imagine a man working in a factory that produces stickers with an inscription: "Don't blame me: I voted for Trump".

[12] This may be treated as a different formulation of the so-called semantic transparency principle. According to this principle, derived from (Husserl, 2001), the user of a sign is focused on its meaning and not on the sign itself, while using it (for a discussion of the principle, see, e.g., Ossowski, 1926; Koj, 1963).

No matter how many stickers the man produced, we will not count them as expressions of his political opinions. He is just not a referent of "I" printed on these stickers. Therefore, we cannot praise or blame him for the sense expressed by the inscription.

Maybe then the referent of the use of the token of "I" is always its user, as stipulated by (SR-1)? With the user/producer distinction it is not difficult to come up with a counterexample to such interpreted Simple Rule. Consider the following example. Alice and Bob are neighbors. Bob is well known in the community as a reliable member of the Democratic Party, who would never vote for a Republican. Alice wants to play a prank on Bob, so she buys a sticker "Don't blame me: I voted for Trump" and she puts it on Bob's door. The prank may not be very witty, but if it is successful, the readers of the inscription on the sticker will interpret the token of "I" as referring to Bob.

It seems therefore that we have a counterexample for both the user interpretation (SR-1) and the disjunctive interpretation (SR-3) of the Simple Rule. Neither Alice nor Bob are producers of the token of I. Alice is the user, but not the producer, and Bob is neither the user nor the producer of the inscription on the sticker. It is Alice who acted intentionally to express a proposition that Bob voted for Trump. Bob wasn't even aware of her action. But it is Bob, not Alice, to whom the use of the token of "I" refers.

Someone may point out that the use of "I" in the example refers to Bob because the interpreter believes that Bob is the user of the token. Alice's prank consists of pretending that it was Bob who put the sticker on the door. But it is not a plausible response. When a man walks around with a "kick me" sticker on his back, no one thinks that he put the sign by himself. And even in such a case, "me" refers to the man. If Alice's prank is revealed, Bob would say "Someone made a political confession on my behalf" rather than "I told you! I just wasn't the referent of 'I' on the sticker".

It seems that a proper user cannot be ignorant of using a linguistic token. The use of a linguistic token is an intentional action. As such it can impose certain commitments on the agent—the user. The nature of such commitments varies with respect to specific kinds of uses. For example, when a user uses a token of the sentence *p* to make an assertion, she is committed to believing that *p*. When she uses a token of the sentence "Is *q*?" to ask a question, we normally expect that she does not know whether *q*.[13] We cannot expect the fulfillment of such commitments based on unintentional quasi-uses as Bob's in the example above. It just would not be right.

The other response to the example may be that such uses of the tokens of "I" just fail to refer. But it is counterintuitive. These uses are treated by normal, competent users of language as if they referred. They understand such uses perfectly and treat them as meaningful uses of language. I see no other motivation

[13] With the exception of very specific kinds of questions, like rhetoric questions, or questions asked during an oral exam.

for the claim that such uses do not refer, than a desperate attempt to rescue the Simple View.

The reader may notice the striking similarity between the presented counterexamples to SV and the case described by Alan Sidelle in *The Answering Machine Paradox* (1991). In his paper, Sidelle presents the example of an utterance that may be seen as a counterexample to the rule similar to (SR). The rule concerns the semantics of the uses of "now" and may be stated as follows:

(SR-Now) The reference of the use of the token of "now" is always the time of the use.

From such a rule David Kaplan derived a consequence that no utterance of "I am not here now" can be true (Kaplan, 1977, p. 509). Sidelle presents the counterexample to this claim. He describes a situation in which such utterance is intuitively true. It is the context of listening to the recorded message "I am not here now" on an answering machine. Intuitively, when it is listened to by the calling person, it is true if and only if the person who recorded the message is not in the place of its use at the time of the call.

Sidelle proposes to resolve the paradox by postulating the existence of so-called deferred utterances. The utterance is deferred when it is not the case that the utterer is in place of the utterance at the time of making it. In the less ambiguous terminology, which is preferred in this paper, such a phenomenon will be called deferred use. Sidelle compares deferred uses to actions performed at distance, like a bomb detonation. If one places a bomb on a plane, one destroys the plane when the bomb explodes, though one may be thousands of miles away (1991, p. 535). Similarly, deferred uses are actions performed at distances (both spatial and temporal). From the description, it is obvious that the use of "I am not here now" is deferred. It does not make sense to say that it is used while recording since the use of a meaningful expression has a fixed reference. It would have an unpleasant consequence that each use of the sentence would be false, despite being perfectly understandable and intuitively true. Whether the production of the token of "I am not here now" in the paradox is also in some sense deferred is a matter for another discussion.

The similarity between presented counterexamples and the case described by Sidelle can motivate us to introduce a generalized notion of deferred use. Generalized deferred use is a use of a token of a linguistic expression which is a different event from the production of the token. The difference may lie e.g., in the time or place of the use and production. Presented counterexamples to (SR-1), (SR-2), and (SR-3) can be viewed as cases of generalized deferred uses. In the case of the "Don't blame me" sticker, its use by Alice is deferred, because it is different as an event from the production which takes place in a different place and time.

A proponent of SV may be tempted to say that her view properly describes all the standard cases of uses of tokens of "I", while counterexamples, which employ cases of generalized deferred uses, can be treated as deviant cases. A defender of

SV may propose that the Simple Rule is in fact a *ceteris paribus* rule concerning only standard, nondeferred uses. The rule may be formulated in the following way:

(CP-SR) If the use of the token of “I” is not deferred, its reference is always the user of the token.[14]

However, I believe that such a rule is also incorrect. Even though there are cases in which the acts of use and production of the token of “I” are identical, there is only one producer identical with the user of the token, who is not the referent of “I”. In short, there are also counterexamples to (CP-SR). For example, consider a puppeteer who is animating a puppet on a stage. She may speak on behalf of an animated puppet followed by a puppet villain and say “I’m in great danger!”. In this case, the token of the sentence “I’m in great danger!” is produced by the puppeteer, who is also at the same time using it to express the proposition that the puppet is in great danger.[15]

One may refute the counterexample, saying that it is not a normal conversational context. The puppeteer is behaving just as she is communicating something, but she is not. She just simulates an utterance of a fictional character. This counterexample inherits all the controversies concerning the use of language in fictional discourse. But a similar example may be given in a nonfictional conversational context. Consider an utterance made by Alice the ventriloquist, who wants to speak on behalf of her friend Bob. Similarly to the case of the sticker prank, we may imagine that Alice wants to make a joke on Bob at a meeting of the teaching staff at the university they both work at. Alice, using her ventriloquist skills, may utter the token of the sentence “I think that the Chancellor is a fool” on behalf of Bob (cf. Corazza et al., 2001, p. 15). In this case, she is the producer of the token of this sentence, and she uses it to communicate something about Bob. Moreover, the use and production are the same physical event, hence it is not a deferred use of the sentence. But the referent of “I” is Bob, not Alice.

I believe that the proposed counterexamples show quite convincingly that the Simple View is not the right stance in the debate on the semantics of “I”, and, more broadly, on the semantics of so-called pure indexicals. This is, however, nothing more than a negative result. I believe that it can and should be taken into consideration by philosophers who want to describe the semantics of indexical expressions, but it does not itself endorse any view on that matter. What it does show is that any theory of indexicals should very carefully take into account the distinction between use and production in proposing and assessing semantic theories.

[14] Of course, it is equivalent to the rule stipulating that the referent of a nondeferred use is its producer since, by definition, the user is identical to the producer in such cases.

[15] Or the fictional character represented by the puppet.

5. Conclusion

In this paper, I argued against the claim that the uses of tokens of the word "I" refer in virtue of the simple fact about the context, namely the fact about who the agent of the context is. I showed three possible interpretations of such a claim: the first that uses of "I" always refer to its producer, the second that they always refer to the user of the token, and the third stating that it refers to either of them. I proposed counterexamples to all interpretations of the thesis and also to its *ceteris paribus* version. These examples show that we have to abandon the Simple View about the semantics of "I", and possibly, about all of the so-called pure indexicals. The paper does not argue in favor of any alternative view about the semantics of "I", but I think it shows that any such semantics has to take into account the distinction between the user and the producer of linguistic tokens.

REFERENCES

Barwise, J., Perry, J. (1981). Situations and Attitudes. *Journal of Philosophy, 78*, 668–691.

Briciu, A. (2017). Idexicals in Remote Utterances. *Philosophia*. doi:10.1007/s11406-017-9909-x

Brinck, I. (1997). *The Indexical 'I'. The First Person in Thought and Language*. Dordrecht: Springer.

Campbell, J. (1994). *Past, Space and Self.* Cambridge, Mass.: MIT Press.

Cappelen, H. (1999). Intentions in Words. Nous, *33*(1), 92–102.

Ciecierski, T. (2021). Indexicality, Meaning, Use. *Semiotica*, *238*, 73–89.

Corazza, E., Fish, E., Gorvett, J. (2002). Who is I? *Philosophical Studies*, *107*, 1–21.

de Gaynesford, M. (2006). *I: The Meaning of the First Person Term*. Oxford: Oxford University Press.

Hawthorne, J. Lepore, E. (2011). On Words. *The Journal of Philosophy*, *108*(9), 447–485.

Husserl E. (2001). *Logical Investigations* (Volume II). London and New York: Routledge.

Kaplan, D. (1979). On the Logic of Demonstratives. *Journal of Philosophical Logic, VIII*, 81–98.

Kaplan, D. (1989a). Demonstratives. In: J. Almog, J. Perry, H. Wettstein (Eds.), *Themes from Kaplan* (pp. 481–563). New York: Oxford University Press.

Kaplan, D. (1989b). Afterthoughts. In: J. Almog, J. Perry, H. Wettstein (Eds.), *Themes from Kaplan* (pp. 565–614). New York: Oxford University Press.

Kaplan, D. (1990). Words. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *64*, 93–119.

Kaplan, D. (2011). Words on Words. *The Journal of Philosophy*, *108*(9), 504–529.

Katz, J. (2000). *Realistic Rationalism*. Cambridge, Mass.: MIT Press.

Koj L. (1963). Zasada przezroczystości a antynomie semantyczne. *Studia Logica: An International Journal for Symbolic Logic*, *14*, 227–254.

Lewis, D. (1970). General Semantics. *Synthese*, *22*(1/2), 18–67.

Ossowski, S. (1926). Analiza pojęcia znaku. *Przegląd filozoficzny*, *1–2*.

Pelc, J. (1971). *Studies in Functional Logical Semiotics of Natural Language*. The Hague: Mouton.

Penco, C. (2021). Indexicals and Essential Demonstrations. *Semiotica, 240*, 261–284.

Predelli, S. (1998). 'I Am Not Here Now'. *Analysis, 58*(2), 107–115.

Sidelle, A. (1991). The Answering Machine Paradox. *Canadian Journal of Philosophy, 21*(4), 525–539.

Strawson, P. F. (1950). On Referring. *Mind*, *59*, 320–344.

Vision, G. (1985). I Am Here Now. *Analysis, 45*(4), 198–199.

Wetzel, L. (2009). *Types and Tokens. On Abstract Objects*. Cambridge, Mass.: MIT Press.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXV, nr 1 (2021), s. 107–118
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxv1.07

Review

ANTONINA JAMROZIK [*]

REVIEW OF PAWEŁ GRABARCZYK'S *DIRECTIVAL THEORY OF MEANING: FROM SYNTAX AND PRAGMATICS TO NARROW LINGUISTIC CONTENT*

SUMMARY: This paper is a review of Paweł Grabarczyk's latest book, *Directival Theory of Meaning: From Syntax and Pragmatics to Narrow Linguistic Content* (2019). I focus mostly on two concepts constitutive for the directival theory of meaning—that of linguistic trial and that of meaning directive. These two concepts, while ingeniously developed by Grabarczyk, are not free of problems and somewhat controversial assumptions. I start with describing the basis of Grabarczyk's proposal, as well as of the historical background from which it originated. Then, I move on to the analysis of the notion of linguistic trial. After that I focus on the concept of meaning directive, criticising certain assumptions that come with it. The conclusion is that while Grabarczyk's version of the directival theory of meaning is an interesting proposal, most of its shortcomings stem from the fact that for a theory that is supposed to work well on natural languages, too many examples pertain to artificial languages. Until an analysis of a natural language in the style of the directival theory of meaning is conducted, it is not possible to properly judge the value of this theory.

KEYWORDS: directival theory of meaning, inferentialism, holism, molecularism, compositionality.

Introduction

In his latest book, *Directival Theory of Meaning. From Syntax and Pragmatics to Narrow Linguistic Content*, Paweł Grabarczyk undertakes an ambitious goal of resurrecting and reformulating a semantic theory created by Kazimierz

[*] University of Warsaw, Faculty of Philosophy. E-mail: a.jamrozik@uw.edu.pl. ORCID: 0000-0001-7717-0591.

Ajdukiewicz (1931). This goal is ambitious for several reasons. First of all, the original version of this theory was developed by Ajdukiewicz in the 1930s, against a vastly different philosophical background. Furthermore, its development was halted by a counterexample provided by Alfred Tarski and Ajdukiewicz himself abandoned the theory. The result is that philosophers of language, during the development of this discipline, have not really engaged in any dialogue with the directival theory of meaning.[1] This means that both the background philosophical assumptions of the theory, and the language in which it was formulated are in need of examination and reformulation. Second, the theory is molecularist. Both molecularist and holistic semantic theories are under significant scrutiny since Fodor and Lepore's argumentation against them (Fodor, Lepore, 1991; 1992). Lastly, the theory is non-referential and rather humble in its explanatory aims, which can, and according to Grabarczyk should, be read as its advantage, but which also calls to question whether it would not be more beneficial to adopt a theory that could explain a more broad range of phenomena related to natural language.

The aim of this paper is to review the ideas put forward by Grabarczyk in his book. Due to the fact that a vast range of subject is covered in said book, including the comparison of the directival theory of meaning (in both its original and new version) to other semantic theories, I will not concern myself with every claim that Grabarczyk makes, but rather limit myself to the most substantial ones or the ones that might seem controversial. However, in order to better explain said substantiality and controversy, I will start with shortly presenting the main tenants of the Ajdukiewicz's version of the directival theory of meaning and the weak points that Grabarczyk identifies within it, and later tries to amend in the new version of the directival theory of meaning (hereafter nDTM, following the author I will also use DMT when talking about the directival theory of meaning in general). The plan for the remainder of the paper is the following: After a brief description of the DTM I will focus on the notion of semantic trials and the assumptions that Grabarczyk makes with regards to them. It is one of the two most important notions in the DTM, so it should come as no surprise that afterwards I will turn to the other one—that of meaning directive. I will consider this notion with regards to the structure of meaning directives, meaning I will concern myself with both how are they structured internally and the properties of the structure the set of them once they are collected. I will finish with some general remarks about the nDTM, its scope, as well as advantages and disadvantages presented in the book.

[1] At least on the face of it; as Grabarczyk notes, the semantic theories of Wilfrid Sellars and Robert Brandom both bear striking similarity to that Ajdukiewicz. While it is possible that some of this similarity boils down to the fact that all those theories are non-atomistic, the degree of similarity is still quite striking.

2. Directival Theory of Meaning—The Basics and the Problems With Them

Grabarczyk himself starts his book by presenting the obvious starting point for the development of his own theory—the DTM as it was formulated by Ajdukiewicz. It is worth noting that already at this stage he modernises the vocabulary and the formal apparatus that the theory was originally expressed with. His discussion focuses mostly on the formal layout of the theory, as its motivations are of secondary importance in the modifications.[2] The key concept to the DTM is that of meaning directive. In order to understand what meaning directives are, it is prudent to start with the notion of semantic trial. As Grabarczyk points out, the assumption about language that can be considered to be foundational for DTM is that there is a set of sentences, platitudes, that a person must accept (either unconditionally or under certain conditions) in order to be treated as a member of a linguistic community by said community. Furthermore, every expression of a given language appears in some of the sentences belonging to this set. A semantic trial is a situation in which a person's belonging to a given linguistic community is tested by checking if she accepts a sentence belonging to this set. As said above, this acceptance might be conditional or unconditional. To follow Grabarczyk's examples—if a person's linguistic behaviour is such that people in the community are suspicious of whether she uses the word "table" in the correct way, for example she claims that tables are really friendly and she enjoys talking to them, they might test her by asking the question "Are tables pieces of furniture?". This is due to the fact that a sentence "Tables are pieces of furniture" belongs to the set of platitudes that have to be unconditionally accepted by every member of the linguistic community. If someone is suspicious of whether a person uses the word "cold" correctly, they might hand her an ice cube and ask the question "Is this cold?". This, one the other hand, is due to the fact that a sentence "This is cold" belongs to the set of sentences that have to be conditionally accepted by the members of linguistic community, under the condition that they are presented with a cold object. The set of meaning directives is defined as a set of rules that specify under what conditions a person has to accept what sentence in order to be considered a member of a given linguistic community. What is important to know is how the semantic trials function and what is the structure of meaning directives, as this is the heart of DTM.

Having explained the concept of meaning directives, it is possible to define the notions of meaning, synonymy and translation. Meaning is relativised to the structure of the set of meaning directives[3] and is defined as an ordered pair

[2] Grabarczyk does in fact provide an extensive analysis of the philosophical motivations and the background against which DTM was created, however, since he rejects most of the assumptions and motivations endorsed by Ajdukiewicz in creating his version of DTM, I will omit this analysis in my review.

[3] According to Grabarczyk, the best way to represent this structure is in a table which contains the representation of a situation in which a given sentence has to be accepted, said sen-

whose first element is said structure and second element—the set of places that are occupied by a given expression in this structure. Synonymy is classically understood as sameness of meaning of expression, so its definition should not be surprising—two expressions are considered to be synonymous if they are interchangeable within the set of meaning directives without changing this set. The definition of translation is perhaps most puzzling, due to its rigidity. Two languages are considered to be translatable into one another if it is possible to structure the set of meaning directives of the two in the same way. The translation of a given expression in one language is considered to be an expression in the other language that has the same distribution in said structure. As I have mentioned, this definition might seem overly rigid. It becomes clear why it is so when we realise that Ajdukiewicz meant for his theory to apply only to closed languages, i.e., languages that contain every possible meaning—no new meaning can be added to them on pain of generating an inconsistency within the language. It is one of the shortcomings of the original DTM rightfully noted by Grabarczyk. His solution is to simply ditch the assumption that DTM is only suited to deal with closed languages. Let us now take a look at other weak points of the original DTM that Grabarczyk identifies and his solutions to them.

Another problematic assumption that Ajdukiewicz makes has to do with the fact that his theory was in fact created in order to give more gravity to his views about philosophy of science, namely the position of radical contextualism. Since Grabarczyk's goal is to reformulate DTM in such a way that it can be useful for the analysis of natural languages, here too he simply abandons this assumption and the consequences it has for the DTM. However, there is one preconception of Ajdukiewicz that does not seem so easy to deal away with. It has to do with one specific kind of directives, empirical directives. The part of the directive that has to do with the circumstances under which a person is required to accept given sentence contains an empirical part. Recall the example with the sentence "This is cold" and presenting one an ice cube. The directive "When presented with an ice cube, accept the sentence 'This is cold'". contains a part which describes an experience of being presented an ice cube. Grabarczyk notes that Ajdukiewicz hesitates as to what language to choose for the description of this empirical component of certain directives, leaning towards the psychological notions, such as motive. And regardless of the choice of the language, there seems to be no way to avoid DTM's commitment to some theory of either mind or external world. Grabarczyk, however, seems to find a way to do so. As any other part of meaning directive, the part of each empirical directive that pertains to the subject's experience occupies a certain place in the structure of all the directives. Therefore, Grabarczyk notes, we can base the identity conditions for the empirical parts of meaning directives on this structure, and look at them in a purely functional way—the experience *x* is the experience that has this-and-this distribu-

tence, and each of the parts of this sentence. However this is not the only possible way of structuring meaning directives, so here I will present it in full.

tion in the structure of the directives. This choice of identity condition does not preclude any theory of mind nor external reality; it could be compatible with any approach to those questions. This is in fact characteristic of the author's approach, as what he seems to be doing is ridding the DTM of the majority of ontological or other philosophical assumptions. It would seem as he wishes to present it as a "pure semantics", which has a very limited scope, and relegate any other job to other philosophical and scientific theories. Hence, the requirement that nDTM is neutral in many aspects is crucial, for only then such relegation is possible.

The last fault of Ajdukiewicz's theory of meaning that I want to mention here has to do with the implementation of his theory to natural languages. The first problem with it is connected to what is mentioned above, namely the assumption that DTM is only suited for the analysis of closed languages. Clearly, no natural language can be considered to be a closed language. Ajdukiewicz circumvents this by claiming that every language is some stage of development of closed language and that one can assume that every language will eventually become closed. Grabarczyk abandons this assumption altogether. A more practical worry concerns the following questions: how are semantic trials to be recognised? And how is the linguist to proceed in order to create a DTM-style theory of a given natural language? The answers that the author gives to these questions is neither simple nor uncontroversial, so I will devote the next section to the analysis of them.

3. The Status and Role of Semantic Trials

The process of discerning semantic trials among other linguistic behaviours and collecting meaning directives is dubbed the "pragmatic part" of the theory, as opposed to the "syntactic part", which consists in parsing the sentences present in the meaning directives and structuring the set of meaning directives itself. Grabarczyk remarks that there is little to none said about the pragmatic part in the works of Ajdukiewicz, and since his goal is to create a theory that can be actually applied to natural languages, he has to fill this void. Let us now look critically at the solutions that he proposes.

First, there is the problem of collecting meaning directives. The author of nDTM is adamant that his theory is not a theory of radical translation, as for a linguist to be able to detect semantic trials she has to have the ability to recognise the semantic trials and correctly judge whether the person undergoing the trial succeeds, i.e., she has to be able to distinguish the acceptance of a sentence from a rejection of a sentence. However, this is not enough—she also has to be able to discern semantic trials from all other sorts of linguistic behaviour. Grabarczyk is conscious that the criteria of identity of semantic trials are by no means obvious. One cannot claim that semantic trials consist of asking platitudinal questions, for it is not clear if certain sentences figure in meaning directives because they are platitudes or if certain sentences are platitudes because they figure in meaning directives. Grabarczyk proposes two features that are supposed to be distinguishing of semantic trials—the use of semantic vocabulary: words such as “meaning”, “reference”, “sense”, etc., and the fact that a person failing a semantic trial is not treated seriously, that her statements are regarded as nonsensical, and that her acceptances or rejections of given sentences cannot be treated as basis for predicting her future behaviour. The latter requirement is designed to capture the difference between semantic trials and other situations in which a person does not conform to the linguistic norms of acceptance of certain sentences. Grabarczyk provides an example of a person failing to accept the sentence “Do not smoke in the mining shaft”. According to him, when a person fails to accept this sentence, this results in them being prohibited from entering the shaft, meaning that their future behaviour is predicted on the rejection of said sentence. Failing a semantic trial does not have such consequences.

Grabarczyk is conscious of the fact that it this is not enough a requirement, hence claiming that the use of semantic vocabulary is another marker of semantic trial. However, it is rather easy to imagine a situation which fulfils both of the requirements and yet intuitively it is not clear at all whether it should count as semantic trial. It is important to remark that nowhere in his book does Grabarczyk claim that linguistic behaviour is substantially different from any other sort of human behaviour. So, when he talks about the acceptance or rejection of given sentence as not being able to provide basis for prediction of future behaviour of a given person as a mark of semantic trial, this extends to linguistic behaviour. Let us now imagine a scenario in which language user suspects that the person she is talking to uses the word “green” in bizarre fashion, raising her suspicion as to whether the person she is talking to is a competent language user. She sets up a semantic trial by saying “I’m not sure we mean the same thing by the word ‘green’”, fulfilling one of the requirements for semantic trial. Further, she asks the person, pointing towards a patch of grass “Is this green?”. The person rejects this sentence. So far, it seems like a perfect example of a semantic trial. However, the language user might find it puzzling why the person she was talking to was perfectly able to communicate all the thoughts and that the only problem appeared with regards to the word “green”. She might further test this person by pointing to a red bench and asking “Is this green?”. Suppose the person accepts this sentence. She might then assume that, for whatever reason, the person uses the word “red” when people normally use the word “green” and vice versa. She might test this suspicion and come to the conclusion that it is true. This is the only bizarre thing in the idiolect of this person, so it does not preclude communication. Therefore, the prediction of the future linguistic behaviour might be drawn from it—the language user simply has to assume that when the person says “green” they mean what she means by the word “red” and vice versa.[4]

[4] This is perhaps reminiscent of Davidsonian radical interpretation, and understandably so, but it is important to remember that unlike Davidson, Grabarczyk does not make any assumptions about the cognitive layout of the language users nor about their psychological preconceptions. The situation described here is to be read in this way, i.e., as a third-person perspective account of what might happen after a person seemingly fails a semantic trial.

If this example seems too far-fetched, consider a community of specialists existing within a certain language. It is possible that different meaning directives set the boundaries of the language when it is spoken among those specialists, and different ones when it is spoken to a person of which they thought that she belonged to the specialists but turned out not to. In the latter case the specialists might simply adjust the language they use when they realise that the person they are speaking to is not a specialist herself. Of course one might in turn claim that the act of realisation that she is not a specialist is actually the same as denouncing her as a non-member of a given linguistic community. However, this would lead to the conclusion that each of the English-speaking specialist group actually does not speak English among themselves but rather different languages, physicist-English, electrician-English, English-teacher-English, and so on. While one might bite the bullet and say that it is in fact the case that these are all different, albeit similar, languages, such statement seems quite counterintuitive. It might also lead us down a slippery slope towards the solipsistic claim that every person speaks slightly different language, even in they are similar to each other. This claim, apart from also being rather counterintuitive, contradicts one of the tenants of the DTM, which is that it is a theory of environmentally narrow, but socially broad meaning.

Last point pertinent to this matter that I want to touch on here is the question of how the notion of metalinguistic negotiation (Plunkett, 2015) relates to that of semantic trial. According to Plunkett, "A metalinguistic negotiation is a metalinguistic dispute that concerns a normative issue about what a word should mean, or, similarly, about how it should be used, rather than the descriptive issue about what it does mean" (Plunkett, 2015, p. 828). From the examples provided above it should be clear that the same words uttered under exactly the same situations can serve as both semantic trial and a start of metalinguistic negotiation. The effects of the two are different, but the similarities between them are not coincidental—since meaning directives determine the boundaries of the language, it would seem likely that metalinguistic negotiation is one of the ways to change these boundaries. However, since Grabarczyk wants nDTM to be a theory that can actually be used to analyse natural languages, I believe that he needs a more clear-cut distinction between semantic trials and other linguistic behaviours in order for the nDTM to be useful in this regard.

4. The Structure of Meaning Directives—Inside and Out

Meaning directives lie at the heart of DTM, as the name of the theory suggests. They are constitutive of the boundaries of language and allow to distinguish meaningful discourse from mere gibberish. According to Grabarczyk, creating this distinction is the primary task of any semantic theory, so it is no wonder that he looks at the notion of meaning directive with great scrutiny. He distinguishes four kinds of meaning directives, adding one to the list proposed by

Ajdukiewicz, who only defines three.[5] First, there are empirical directives, which require language users to accept a given sentence provided they are in certain internal state (described functionally). Second, there are inferential directives,[6] in which the language user is expected to accept a sentence provided that she accepts some another sentence or sequence of sentences. Third, there are axiomatic directives, according to which a language user is to accept a given sentence under any circumstances. Finally, and these are Grabarczyk's own addition, there are promotive directives, which require the language user to perform some action, understood as bodily movement, upon encountering some sentence or a sequence of sentences. The example of such directive could be a situation when a language wants to check if the person she is taking to understands the one is required to stop after hearing the command "Stop!". Of course Grabarczyk is aware that not every instance of uttering such command should count as a semantic trial. However, the way he deals with this issue is problematic, as he claims that "The point here is that once the user recognizes that she is to take a semantic trial and accepts the command, she is expected to act in a certain way" (Grabarczyk, 2019, p. 160). The requirement that a language user is supposed to recognise given situation as a semantic trial seems to be not only inexistent for other directives but also in contradiction to Grabarczyk's insistence that language users do not have to understand directives, they only have to conform to them. One might say that the notion of semantic trial and that of meaning directive, although interconnected, can be understood independently, but such claim would warrant further explanation and evidence. For a theory that strives not to assume anything about the cognitive structure of language users, the requirement that the tested person should recognise the situation she finds herself in seems really strong, especially if other directives work just as well without it. This begs the question what is so special about promotive directives. Well, the biggest difference between them and all the others is that the other ones require the tested person to either accept or reject certain statement, while promotive directives require the tested person to perform much wider class of actions, non-linguistic ones to that. The differences between promotive directives and all the other ones seem to be quite substantial, and I believe that the promotive directives either require another definition or this notion should be abandoned completely. This is however a minor point, since, as it already has been mentioned, the choice of directives depends on the decision of a researcher.

Let us now turn to the question of the structure of the meaning directives. It is of utmost importance to the nDTM, as the notions of meaning, synonymy and translation are defined in relation to this structure. There are multiple ways of structuring meaning directives, although it is clear that the structures have to

[5] Both of them, however, claim that this list is not exhaustive; Grabarczyk remarks that the choice of kinds of directives suitable for an analysis of a given language is also a matter of empirical investigation.

[6] Ajdukiewicz calls these directives deductive, Grabarczyk changes the name for the sake of clarity, but the content of a directive remains the same as in Ajdukiewicz's work.

fulfil certain requirements—the directives in them have to be adequately parsed and cannot be ordered. If the former was not fulfilled then it would not be possible to structurally identify the meanings of single words, and if the latter was not fulfilled, even the slightest change in the order of the directives in the structure in one of the two fully mutually translatable languages would render them untranslatable or only partially translatable. Since the meaning of an expression is determined by its place in the structure of the (parsed) meaning directives, the theory has a strong holistic component.[7] For the reminder of this section, let us consider how it bears against the objections against holism—the problem of compositionality, and against molecularism—the problem of analytic sentences, starting with the latter.

The problem of analytic sentences in molecularist theorists was put forward by Fodor and Lepore (1992). It boils down to the following claims: First, molecularist theories posit that meaning of a given expression is dependent only on a subset of the meanings of other expressions, not their entirety. Second, there is a way to distinguish the set of expressions (sentences) that are meaning-constitutive. From this they assume that the only such set could be the set of analytic sentences. But, as they claim, since the Quinean critique of the analytic-synthetic distinction (Quine, 1953), one cannot reasonably use it in the theory of meaning. Hence, since molecularist theories are based on this distinction, they are unwarranted without it. Grabarczyk claims that this objection is applicable to Ajdukiewicz's version of the DTM, however the nDTM escapes it. If one was to take "analytic" to mean "true in virtue of language rules" then, since nDTM does not assume any conception of truth, it could not apply to the sentences enclosed in the meaning directives. In fact, as Grabarczyk points out, there might be sentences which are plainly false but nevertheless are enclosed in axiomatic directives in a given linguistic community. The only thing that the nDTM can tell us about truth is the meaning of the predicate "is true" in a given linguistic community, as it is distributed in the structure of the directives. Moreover, as the author claims, even if we adopt a meta perspective on the language, the notion of truth is still only relativised to a given language, so even if it is so that according to the directives of this language if one accepts the sentence *p* one should also accept the sentence "*p* is true", we are still talking about the notion of truth as relativised to this language. I believe this line of argumentation to be correct, however, I think that there is another possible way to look at the meta perspective which is worth considering In principle, it is perfectly possible, and perhaps even favourable, to describe the directives in the metalanguage not as input-output scenarios, but rather in sentential form. This might be especially useful in the

[7] Thorough his book, Grabarczyk calls his theory molecularist, rather than holist, but sometimes writes holist/molecularist. Molecularism is considered a more moderate version of holism, so most of the objections against it apply to fully-fledged holism as well. Moreover, under certain definitions of holism, his theory could be considered holistic, as although the meanings of expressions are not determined by the totality of the expressions in a given language, the meanings themselves are interdependent.

early stages of collecting the meaning directives for a given language, as it would simplify the description of the parts of the directives.[8] This way, we could say that sentence such as "When having an experience of a cold object, one has to accept a sentence 'this is cold'" or "Under any circumstances one has to accept the sentence 'Chairs are pieces of furniture'" are descriptions of the meaning directives in a (primitive) metalanguage. Are these sentences analytic in the metalanguage? Since in this language we can, in principle, speak of truth per se, it is an option worth considering, however here too a lot seems to depend on the choice of metalanguage by the researcher.

Let us now turn to perhaps the most famous objection against holistic semantic theories, first put forward by Fodor and Lepore as a counterargument to the conceptual role semantics (Fodor, Lepore, 1991), and later developed in their book (Fodor, Lepore, 1992). According to semantic holism, the meaning of an expression is determined by its relation to other expressions. According to the principle of compositionality, the meaning of a complex expression is determined by the meaning of their parts along with the way they are composed. On the face of it the two seem incompatible, although there have been attempts to reconcile them (e.g., Block, 1993; McCullagh, 2003). Peter Pagin claims that the conflict between these two theses boils down to the question of priority—in holistic theories, the meaning of the whole comes first, while in accordance to the principle of compositionality, the meaning of simple expressions comes first (Pagin, 1997). In this regard, it would seem that nDTM endorses the holistic claim, and therefore is incompatible with compositionality. However, Grabarczyk proposes a way out of this conundrum. In order to evaluate it properly, it is necessary to start from explaining what it means for an expression to figure in a meaning directive in essential manner. Grabarczyk defines what does it mean for an expression to figure in a directive in an essential manner in the following way: "An expression figures in a sentence in an inessential manner if it can be replaced in the sentence by any other expression of the same syntactical category without changing the set of directives. Otherwise it figures in the sentence in an essential manner" (Grabarczyk, 2019, p. 170). This notion is important not only in understanding what does it mean for a language user to know the meaning of an expression (she has to conform to the meaning directives in which this expression figures in an essential manner), but also in order to understand Grabarczyk's solution to the problem of compositionality. He rejects the claim that nDTM should account for strong compositionality, understood as providing a way to create novel meaning directives for complex expressions. Instead, he proposes to introduce a different way to generate meanings of complex expression. In this he claims to endorse a weak version of compositionality—a claim that the meaning-securing mechanisms are different for simple expression and different for complex expressions. In order to define the meanings of complex

[8] This is mostly due to the fact that the functional description of the language user's internal states is available only after the directives have been collected and structured.

expressions, Grabarczyk proposes to consider a structure SD, which is a set of directives D plus all the directives resulting from the substitution of variables in the sentences enclosed in the directives in D. This means that in creating SD, one abandons the requirement that the expression in the sentences enclosed in the directives figure in them in an essential manner—SD is a set of all properly built sentences in the language. While this is true for the toy language that Grabarczyk bases his examples on, it might not necessarily be true for a natural language such as English. There is no way to guarantee that every possible grammatical structure of a sentence is exemplified by some sentence in the set of meaning directives, as discovering meaning directives is a matter of empirical investigation. It is in principle possible that it turns out that there are no sentences having a specific structure. I fail to understand how generating the set SD secures the meaning of every complex expression—this, while certainly possible, is dependent upon contingent factors, and compositionality, even in its weak reading, is a necessary feature of a natural language. Moreover, I find the concept of figuring in a directive in an essential matter confusing—an example of what it means to figure in a directive in an inessential manner provided by Grabarczyk is an inferential directive for conjunction—regardless of what are the conjuncts, if one accepts both conjuncts, one has to accept their conjunction as well, regardless of what the conjuncts are. Inessential elements of the sentences can be represented by variables. It does, however, seem hard to implement this rule while collecting meaning directives, and the choice of whether certain expression figures in a sentence enclosed in a directive in an essential or inessential manner—one's rejection of the conjunction of the two accepted conjuncts could stem from the fact that she does not know the meaning of the word "and", or it could stem from the fact that she associates a specific meaning with the two conjuncts—she accepts both of them separately but not in conjunction with each other. The motivations for idea that certain expressions figure in the directives in an essential manner seems clear, but since nDTM is molecularist/holist, the line between inessential and essential manner is not as clear cut as it would seem at first glance.

5. Closing Remarks

As mentioned in the beginning, Grabarczyk sets rather ambitious goals for his book, most of which seem to be met. In this review I drew attention mostly to its controversial fragments, it is however worth remembering that the theory itself is an interesting proposal, especially against the background of existing semantic theories. What is interesting is that most of the features that prompted me to classify this theory as humble in its explanatory aims are also responsible for its uniqueness. In short, Grabarczyk's approach seems to be to abandon the controversial philosophical assumptions of Ajdukiewicz and at the same time preserve most of the features of the theory that were motivated by those assumptions, only motivating them otherwise. The nDTM is non-referential, it focuses on determining the boundaries of language rather than its internal features, it does not assume any sort of cognitive structure of language users, it is socially narrow, it does not fulfil the requirement for strong compositionality, it embraces Tarski's challenge, and finally, it remains agnostic with regards to the definition of truth. Moreover, it is environmentally wide, assumes that syntactic structures are prior to semantic structures. When comparing it with semantic theories with a much wider explanatory aim and much more assumptions, such as that of Quine, Davidson or Sellars, Grabarczyk says that while they may possess many advantages over nDTM, the nDTM trumps them in one regard—those theories are thought experiments and as such are impossible to be implemented in linguistic practice, while the nDTM should be regarded as providing a recipe for the analysis of actual languages—either artificial or natural. The examples provided in the book are of artificial languages or merely fragments of natural languages—no wonder, as providing a nDTM-style analysis of most natural language in existence would be extremely laborious task that would require a lot of field research, psycholinguistic studies, and would take up a lot more space. However, I believe that in order to properly judge the value of nDTM as a recipe for an analysis of a language, it would be highly beneficial to see it in action, i.e., to see how can it be applied to a concrete natural language. I would be extremely curious to see such result and hope that will see them in the future.

REFERENCES

Ajdukiewicz, K. (1931). On the Meaning of Expressions. In J. Giedymin (Ed.), *The Scientific World-Perspective and Other Essays, 1931–1963* (pp. 1–34), Dordrecht: Springer.

Block, N. (1993). Holism, Hyper-Analyticity and Hyper-Compositionality. *Mind and Language*, *8*(1), 1–26.

Fodor, J. A., Lepore, E. (1991). Why Meaning (Probably) Isn't Conceptual Role. *Philosophical Issues*, *3*, 15–35.

Fodor, J. A., Lepore, E. (1992). *Holism: A Shopper's Guide*. Oxford: Blackwell.

Grabarczyk, P. (2019). *Directival Theory of Meaning: From Syntax and Pragmatics to Narrow Linguistic Content*. Cham: Springer.

McCullagh, M. (2003). Do Inferential Roles Compose? *Dialectica, 57*(4), 431–38.

Pagin, P. (1997). Is Compositionality Compatible With Holism? *Mind & Language*, *12*(1), 11–33.

Plunkett, D. (2015). Which Concepts Should We Use? Metalinguistic Negotiations and the Methodology of Philosophy. *Inquiry*, *58*(7–8), 828–874.

Quine, W. V. (1953). The Problem of Meaning in Linguistics. In *From a Logical Point of View* (pp. 47–64). Harvard: Harvard University Press.