

BALKAN JOURNAL OF PHILOSOPHY

Vol. 14, Issue 2, 2022

The issue is sponsored by the Bulgarian National Science Fund, competition “Bulgarian Scientific Periodicals – 2021”, project No: KII-06-HII 3/6 from 15.11.2021

TABLE OF CONTENTS

Articles

Invited paper

Hobos – on the Fragility of Human Diachronic Identity in Whitehead’s Process
Philosophy / 83

Franz Riffert

Invited paper

Truth and Modalities (I) / 103

Fabrice Pataut

A bilateral reduction sentence for modulation / 109

Miguel López-Astorga

Moral Challenges for Bauer’s Project of a Two-level Utilitarian AMA / 115

Silviya Serafimova

Mental representations of political discourse in an authoritarian society: The case of Albania during the implementation of the Chinese Cultural Revolutionary model / 127

Majlinda Bregasi, Albert Bikaj

The criminal responsibility of high-functioning autistic offenders in Croatia / 137

Mladen Bošnjak, Marko Jurjako, and Luca Malatesti

Evidence and Proof of God’s Existence, According To Hegel / 149

Dragoş Popescu

Understanding and Imagination. A Kantian Interpretation of Hegel's "Inverted World" /157
Vlad Bilevsky

Quantum Mechanics, Being and Cognition / 163
Ivan Katzarski

Bootstrap's Monadology. Symmetry and Mirroring Connections between Chew's Bootstrap Theory and Leibniz's Monadology / 173
Ramona Ardelean

Book review

Gunnar Skirbekk. Religion in Modern Society (Religion i moderne samfunn).
Dreyers Forlag Oslo. 2021 / 183
Silvia Serafimova

Röck, Tina (2022) *Dynamic Realism Uncovering the Reality of Becoming through Phenomenology and Process Philosophy*, Edinburgh University Press / 188
Marina Bakalova

Articles

Invited paper

HOBOS* – ON THE FRAGILITY OF HUMAN DIACHRONIC IDENTITY IN WHITEHEAD'S PROCESS PHILOSOPHY

Franz Riffert

Department of Educational Science, Salzburg University,
Erzabt Klotz Straße 1, A-5020 Salzburg, Austria
email: Franz.Riffert@plus.ac.at

Abstract:

After an exposition of the importance of the concept of diachronic identity and a short sketch of the bewildering confusion concerning the meaning of the terms intimately connected to it, an outline of the basic features of Whitehead's process approach is presented in as far as it is relevant for the topic of the human person and identity. In a further step, Whitehead's concept of 'person' as nested in this approach will be discussed and the dilemma between securing human identity on the one side and accounting at the same time for its flexible adaptivity to the environment on the other, will be elaborated. Then the concepts of strict and partial identity (in the sense of equivalence) and the problem of the fragility of human identity, and even its possible loss, and how identity can be secured within this process-organismic conception of 'person' will be dealt with. Finally, the sources of partial diachronic identity within the framework of process philosophy will be discussed.

Key words: Process philosophy, Whitehead, human person, living personal society, (strict and partial) identity, equivalence, identity vs. adaptability, fragility of identity, physical memory, God's consequent nature.

1. Introduction: Exposition of the Problem

Rescher in chapter 6, *Process and Persons*, of his book *Process Metaphysics – An Introduction to Process Philosophy* points out that the application of a metaphysical process approach to anthropological

and psychological topics “*offers a far more promising prospect*” (1996, 105) than do traditional positions nested in the “framework of substance ontology” (1996, 105). And no doubt he is right. But it is no less true that process approaches nevertheless suffer their own immanent systemic problems. While dynamic approaches can better account for the activity and flexibility of persons, i.e. their sensori-motor, emotional and cognitive adaptive activities, it is just as true that the question of how diachronic identity or at least some kind of continuity of the person over time can be secured amid all the environmental changes throws its alarming shadows upon the process philosophic approach.

This problem is not to be underestimated for it seems to be intimately connected with moral, rational and emotional issues. For instance, Madell (1981, 115f) has stressed: “If identity is a ‘fiction’, then I cannot suppose that I should feel any shame for my wrongdoing, or regard myself as responsible for it. Shame, remorse, pride and gratitude are just a few of the moral attitudes which clearly depend on a conception of persons as enduring entities”. Further, delaying any immediate gratification or advantage in favour of a more valuable gratification or advantage in the future would not make sense, for it would be *somebody else* who would profit in the *future* from *my immediate* delay in the *present*. If a person for instance is nothing but a bundle of successive micro-processes then why should I care about the process that just vanished or for the process starting in a moment? On an even more basic level: In what sense could one label a stream of successive (conscious and/or unconscious, physical and/or mental) processes a (unified) ‘person’? Therefore, according

* The Macmillan Dictionary gives the following (American English) definition of ‘Hobo’: “Someone who travels from place to place, because they do not have a home or a job.”

to Rescher, the main obstacle an atomist process approach has to face is the problem of identity: “Personal identity, in particular, meets with grave problems when approached as a construct of discrete elements” (1996, 112).

Before turning to this problem in more detail, it is helpful to first have a look at the general state of affairs in this field. If today the term ‘person’ is mentioned, usually a whole bunch of associations comes to one’s mind – ‘individuality’, ‘self’, ‘character’, ‘subjectivity’, ‘personality’, ‘substance’, ‘selfhood’, ‘I’, ‘singularity’, ‘ego’, ‘subject’, ‘core self’, ‘individual’ – and the meanings of all these concepts are far from clear; even in the scientific and philosophic literature no widespread consensus concerning the use of these terms, let alone generally accepted precise definitions, can be found. Indeed, we see the term ‘substance’ identified with the concept of ‘person’ by some authors, while others conceive of them as contradicting opposites. For some the concept of ‘person’ embraces the concept of ‘ego’ among other aspects, while for others the two concepts are explicitly identified. In some publications ‘self’ and ‘I’ are used as interchangeable terms, in others they refer to different aspects of the ‘person’ which have to be distinguished very carefully. Between ‘subject’ and ‘person’ one author can find no difference while the other discovers an unbridgeable gap. So David Galin, to mention but one author working in that field, in respect to the term ‘self’ states: “There is much confusion over exactly what *self* means, for ordinary folk and for the academics and professionals who are supposed to be experts on it. Professionals in psychology and its neighbor disciplines (cognitive science, philosophy of mind, psychiatry, behavioral neurology) are no more coherent about self than lay people” (2003, 313). The same holds without much qualification for the rest of the terms mentioned above, so the terminological situation in this field of research must be summarized in the following way: A greater chaos of meaning in this domain of discourse is hardly imaginable!

Rescher’s rough distinction between substance and process approaches, although introducing at least some order, is however still too crude: He correctly draws attention to the fact that under the heading of ‘process philosophy’ there are “distinct approaches to implementing its pivotal idea of pervasiveness and fundamentality of process, ranging from a materialism of physical processes (as with Boscovitch) to speculative idealism of psychic processes (as in some versions of Indian philosophy)” (1996, 8). If this is so – and there seems to be no

doubt about it – this distinction is not able to provide a sufficiently detailed basis for further discussion. In the domain of anthropology and psychology, questions about intentionality, mentality, awareness and self-consciousness are at the very centre of interest. Hence a distinction which does not even touch for instance the antagonism of materialism vs. mentalism (idealism) fails to provide a reliable basis for meaningful discussions. This is one reason why this paper will restrict itself to the discussion of one concrete process approach, namely Whitehead’s process philosophy of organism. Another reason for choosing Whitehead’s philosophy is Rescher’s judgement that “Whitehead’s process atomism of ultimate micro-constituents (‘actual entities’) creates avoidable difficulties for his metaphysics. Process philosophy is better served by a theory of processual levels that treat processes as existing in an interpenetrating order of levels of scale (much as with waves in water)” (1996, 112). Therefore, Rescher seems to take his refuge in a Hegelian-like hypostatizing of the “synoptic unity of process” (1996, 108), which is to be conceived as a “macroprocess” (1996, 108). But, whether this is the only rational choice is far from clear. It is clear, however, that Whitehead would hardly have accepted Rescher’s suggested solution for he writes: “It is obvious that we must not demand another mentality presiding over these other actual entities (a kind of Uncle Sam, over and above all the US citizens)” (PR 108). And in *Modes of Thoughts* he states in respect to the human person, which he also terms ‘soul’: “For example consider the enduring self-identity of the soul. The soul is *nothing else than the succession of my occasions of experience*, extending from birth to the present moment” (MT 163 italics added). Whitehead goes on to explicate what this means: “Now, at this instant, I am the *complete* person embodying all these occasions” (MT 163 italics added).

But Whitehead’s clear rejection of Rescher’s proposed solution for the problem of identity within his process philosophic account introduces urgency surrounding the question of whether his process account of identity is a plausible one. It will therefore be the purpose of this paper to investigate whether Whitehead’s atomist process approach provides us with a satisfying concept of diachronic personal identity, i.e.: a concept which is at least compatible with basic scientific research (in particular the well-established finding that human beings are able to react very flexibly to the intrusions of external stimuli) and common sense assumptions, for

instance those which underlie our basic ethical and legal demands.

In the following sections I therefore will first give an outline of the basic features of Whitehead's process approach. In a second step Whitehead's concept of 'person' as nested in this approach will be discussed. Then the problem of the fragility of identity and how identity can be secured within this process conception of 'person' will be dealt with.

2. Whitehead on Identity: Inert Substance vs. Relational Process

In order to be able to understand Whitehead's radical processual position it is necessary to turn to what he has called the 'subjectivist principle'. Despite its disastrous implications the subjectivist principle has been held explicitly or implicitly by most philosophers since Descartes. Whitehead states: "The history of modern philosophy is the story of attempts to evade the inflexible consequences of the subjectivist principle, explicitly or implicitly accepted" (PR 157). It rests upon three premises which are rejected by Whitehead altogether:

- (i) The acceptance of the 'substance-quality' concept as expressing the ultimate ontological principle. (ii) The acceptance of Aristotle's definition of a primary substance, as always a subject and never a predicate. (iii) The assumption that the experient subject is a primary substance (PR 157).

Premise (i) states that the basic structure of reality can adequately be accounted for in terms of the metaphysical substance-quality conception. The plausibility which obviously is inherent in this conception according to Whitehead is due to the fact that it mirrors the subject-predicate structure of our ordinary languages. This fact augments the subjectivist principle with an air of almost evidence-like plausibility since we always seem to presuppose substances as designated by grammatical subjects and qualities ascribed to these substances by predicates. Sentences such as 'This stone is grey' thereby unconsciously gain the status of paradigmatic illustrations of basic metaphysical relations (see: PR 78).¹ Whitehead readily accepts the importance of

these conceptions for negotiating everyday situations because they have been developed in these contexts during the evolutionary process and thus we all constantly apply them with great advantage in these contexts. However, he frequently cautions against the uncritical transference of these conceptions into the field of metaphysics. In metaphysics we have to deal with the utmost generalities of reality and the constitution of real objects, not with stones and the ordinary demands of the marketplace, such as counting apples and bolts.

Premise (ii) qualifies the term 'substance'. Aristotle's concept of primary substance (*prote ousia*; see his Met. V, 1017b [24]) is characterized by the fact that primary substances are not predicable of other primary substances; indeed, no primary substance can be predicated of anything else. This implies that there cannot exist any internal relations between such primary substances. It is one of Whitehead's major aims to replace this concept of primary substance with his doctrine of actual entities which by definition do overlap and inhere in one another and are thereby internally related.

According to premise (iii) the subject finally is identified with a primary substance. This has far reaching consequences: If we conceive of the experiencing subject in terms of a primary substance the perceptual (and mental) states accordingly have to be considered attributes or qualities of substances since there is no other way to predicate them of a subject so conceived. (PR 51) Therefore it follows from this premise that a person's mental and psychical (e.g. perceptual) acts cannot represent other persons (i.e. substances). They can only represent their (external) qualities or attributes. Radical empiricists such as Hume (see: 2007 [1739]) therefore have been forced by this situation to deny the very existence of substances: If whatever exists must be perceivable, and if substances cannot be perceived, there is no way out of this conclusion. Transcendentalists such as Kant, leery as they were of abandoning the concept of substance, nevertheless were also compelled to declare them inconceivable. In Kant's terms: while the *phainomenon* (the 'Ding für uns'/the 'thing for us') is conceivable, the *noumenon* (the 'Ding an sich'/the 'thing for itself') is not. The result is a disastrous solipsism. Whitehead tries to escape these unfortunate consequences of the 'subjectivist principle' by substituting the Aristotelian concept of a 'primary substance' with his alternative concept of an 'actual entity'. The 'substance' qua actual entity is the outcome (product) of the (passed) actual entities as well as the result of its own acts of

¹ The same line of argumentation can also be found in Nietzsche's *Über Wahrheit und Lüge im außermoralischen Sinn* [On Truth and Lies in a Nonmoral Sense] (1980/1873, 873-890).

self-determination towards its own relevant future. Indeed, “[i]t is fundamental for the metaphysical doctrine of the philosophy of organism, that the notion of an actual entity as the unchanging subject of change is completely abandoned” (PR 29). The concept of ‘actual entity’ plays a key role in Whitehead’s metaphysics and thus has important implications for Whitehead’s concept of a person; consequently, a short outline of his metaphysical system based on the concept of ‘actual entity’ is necessary in order to avoid grave misunderstandings.

2.1 Whitehead’s metaphysics – short, shorter, shortest

It is important right from the start to draw attention to the fact that Whitehead’s use of the term ‘subject’ is quite unusual. It does not designate the human person – neither (what is usually called) her/his physical nor his/her mental (cognitive) aspect. Whitehead instead gives the term a metaphysical meaning, referring to the internal aspect of the micro-processes (actual entities), which further interact with each other to form more complex systems (societies). Since there is only “one type of actual entities” (PR 19), his metaphysics is a “one-substance cosmology” (PR 19). His metaphysics can therefore be characterized as an ‘atomic one-type-monism’. Everything existing is either an actual entity or an aspect of an actual entity or constituted by actual entities. Consider:

‘Actual entities’ – also termed ‘actual occasions’ – ... [are] the final real things of which the world is made up. There is no going behind actual entities to find anything more real. They differ among themselves: God is an actual entity, and so is the most trivial puff of existence in far off empty space (PR 18).

Actual entities are micro-processes which, immediately after having reached their definite unity (by introducing themselves as data for new developing processes), decay. So, ‘to decay’ does not mean to have vanished, but means to change into the function as datum for newly developing actual entities. An actual entity reaches its individual unity by assimilating (integrating) its predecessors. Whitehead here speaks of the concrescence of an actual entity – its growing together. Apart from its process of concrescence and decay an actual entity does not exist. In its aspect as concrescing entity an actual entity ‘overlaps’ with all its predecessors which form its data; in its aspect as decaying actual entity it overlaps or interpenetrates with the newly developing

actual entities for which it functions as a datum. With this double overlap of an actual entity with its past as well as with its future Whitehead introduces an entirely new metaphysical concept which replaces Aristotle’s concept of primary substances (πρῶται οὐσίαι).

Further, each actual entity is bipolar, consisting of a physical and a mental pole. An actual entity starts its own process of self-creation by physically prehending its predecessors, which Whitehead also calls in its entirety ‘datum’. Physical prehensions are mere passive acts of reception of the datum. The mental pole is constituted by conceptual prehensions which are actively assimilating the ‘material’ by physically grasping it, thereby accommodating these ‘materials’ mutually until a final unity is achieved. When the definite unity of the actual entity is reached – Whitehead terms this aspect of an actual entity the superject – the actual entity decays and functions among all the other decaying actual entities as the (partial) datum of newly developing actual entities. In this way actual entities constitute a pulsating dynamic web or grid of internally related momentary micro-processes which build up the universe and push it forward into adventurous novelty.

An actual entity reaches its definitive unity by realizing a definitive pattern. These patterns are realized as ‘eternal objects’. Although Whitehead has borrowed the term ‘eternal object’ from Plato it is important to be aware of the differences between Platonic Forms and Whiteheadian eternal objects: contrary to Platonic Forms, eternal objects have no existence in themselves but for the existence of actual entities. In order to have existence, eternal objects must be aspects or components of actual entities.

All eternal objects are included in the mental pole of one particular actual entity; this one actual entity Whitehead terms God². God provides a sample of graded eternal objects for each single actual entity as possible patterns of unity for its process of concrescence. Each emerging actual entity reacts to these graded possibilities by deciding for itself which one to choose as its subjective aim. This means that an actual entity is never completely determined – neither by its given data, nor by God’s offered initial aims. God’s exercise of influence is therefore not to be understood as coercion but is more correctly described in terms of ‘luring’, ‘persuading’ or ‘convincing’ which leaves room for the

² God does not create these eternal objects; they instead constitute his mental pole or ‘primordial nature’.

actual entity to execute a conditioned free decision (see: PR 344). The datum and God's graded presentation of initial aims together form the *conditio sine qua non* of the newly emerging actual entities. In order to arrive at a sufficient condition of its own being, each actual entity has to actively choose one special initial aim, thereby making it its own subjective aim. This subjective aim guides the actual entity's own process of concrescence. In this qualified sense each actual entity is a (partial) '*causa sui*'. The subjective aim is at the same time directed at the immediate realization of a value in the concressing actual entity and at its relevant future as a particular datum in future concressing actual entities. "This double aim – at the *immediate* present and the relevant future – is less divided than appears on the surface. For the determination of the *relevant* future, and the *anticipatory* feeling respecting provision for its grade of intensity, are elements affecting the immediate complex feeling" (PR 27).

Usually the decision for one specific subjective aim is executed blindly (i.e. unconsciously) in the form of an aesthetic-mental reaction towards the passively grasped datum and an unconsciously-vague prehended relevant future. The amount of self-determining activity is therefore correspondingly low in such cases; only in those highly developed, complex actual entities which have reached consciousness does the mental pole gain enough importance to play a decisive role in the (self-)formation of the actual entity in question. This aspect of the conceptual initiative of the mental pole transforms an actual entity from a purely aesthetic subject into one that is *also* ethical. The ethical subject, due to the conceptual power of imagination in its mental pole, can extend the realm of its anticipated future, which opens a greater array of different perspectives according to which the person can develop his/her life. Morality is thus intimately connected with consciousness. "The greater part of morality hinges on the determination of relevance in the future. The relevant future consists of those elements in the *anticipated future* which are felt with effective intensity by the present subject by reason of the real potentiality for them to be derived from itself" (PR 27 *italics added*). According to Whitehead, whether an action is to be valued as good or bad depends on whether and/or to what extent the relevant future has been taken into consideration. If only the immediate self-benefit has been taken into account the activity is to be considered bad – even if it may lead to a complex satisfaction in the present: "Evil, triumphant in its enjoyment, is so far good in itself; but

beyond itself it is evil in its character of a destructive agent among things greater than itself" (RM 83). Immoral activities can lead to or be accompanied by complex aesthetic contrasts (on this point Whitehead insists). What is bad about them is that they have been realized despite the actor's awareness that these activities do not contribute to the improvement of *other* actual entities in the future, and may even be highly impoverishing or destructive to them.

The development of consciousness in Whitehead's approach is also intimately connected to his concept of 'society', since societies provide the necessary conditions for the development of actual entities that may be capable of reaching consciousness. The universe is the totality of actual entities whether organized in societies or not. 'Societies' are groups or classes of actual entities among which a certain type of social order holds. A certain type of social order is established by the fact that a specific defining characteristic (a definite pattern) as a common element in each of the members (actual entities) of the class is realized. As long as this defining characteristic is successively reactivated, i.e. inherited, in future actual entities, the society will continue to exist as an enduring object with some kind of identity over time (see for instance: AI 203). There are different kinds of societies. Whitehead distinguishes between 'personal societies' and 'corpuscular societies'. A society is called 'personal' if it solely consists of a serial sequence of actual entities which follow each other and pass on a certain characteristic pattern to the next one in the row. It therefore can also be called an 'enduring object' (PR 109) which is "purely temporal, meaning that there is only one actual occasion at a time" (Griffin 1998, 185). Let A, B and C be actual entities which are successive members of a serially ordered society. In this case B inherits a defining characteristic from A by directly overlapping or interpenetrating with A. C inherits its defining characteristic directly from B by overlapping with it and *indirectly* from A by virtue of B's previous partial inclusion of A.

Corpuscular societies on the other hand consist of a combination (web) of several distinct strands of such personal societies. Therefore they can be termed "*spatiotemporal*" (Griffin, 1998, 185) and can also form the phenomenal 'things' that are available to ordinary sensory experience. The combination of temporal societies into a higher-order society implies that all of these strands of personal societies in order to be parts of the one corpuscular society must in addition to their own defining characteristic also express a common pattern which de-

finer the corpuscular society as such. Because of this, there may be different grades of corpuscularity: “A society may be more or less corpuscular, according to the relative importance of the defining characteristic of the various enduring objects compared to that of the defining characteristic of the whole corpuscular nexus” (PR 35).

Furthermore two types of corpuscular societies have to be distinguished. Griffin (1998, 186) calls them “aggregational” and “compound” (following Hartshorne 1972a, 57). These two types differ in their organization and therefore also in their way of functioning. An aggregational society (e.g. a stone) is characterized by a defining pattern which is expressed by all its constituting members. (Note how this differs from the ordinary meaning of ‘aggregate’ which may be exemplified by referring to (a more or less chaotic) pile of wood.) In contrast to a compound society, it lacks any central direction of activity. In compound societies *one* of the serially ordered personal societies that the compound society is composed of dominates (at least for certain amounts of time) the other societies. This dominant society contributes to the whole corpuscular society a direction in experience and activity. So the whole compound corpuscular society (also termed compound *individual*) experiences and acts as if it were a single directed unity. The border between aggregational and compound societies is not clear-cut and “exactly which enduring things should actually be considered compound individuals, rather than mere aggregational societies, is an empirical question, to be settled in terms of whether behavior seems to require a central agent with an element of spontaneity or self-determination” (Griffin 1998, 186).

To summarize: Reality according to Whitehead is constituted by one type of microscopic event. Such actual entities form different types of societies, as well as societies of societies, societies of societies of societies, etc. A corpuscular society, for example, can nest in many other corpuscular societies that exist at different levels and exhibit varying degrees of complexity. An animal body is an example of such multilevel societies.

2.2 Whitehead’s concept of person: the living personal society

A human individual is a highly complex arrangement of different types of societies: elementary particles, atoms, molecules, cells, organs and bones. (AI 206). The human body consists of a complex arrangement of (serially organized personal and corpuscular) societies. These societies and sub-

societies are, to a certain extent and at least partially, hierarchically ordered. In humans and higher-order animals, and to a lesser degree in lower lifeforms such as jellyfish or worms (see: PR 108), the body in general and the (central) nervous system in particular are hierarchically arranged so as to make possible the emergence of a dominating or ‘regnant’ serially ordered personal society.³ This regnant society may be conceived of as a thread or sequence of actual occasions which reactivate a characteristic pattern that is thereby common to all the members of this society and that defines them as members of this society. Whitehead calls this common pattern the “defining characteristic” (PR 107) of a society. “The self-identity of a society is founded upon the self-identity of its defining characteristic” (AI 204).⁴ Or in more detail:

The one individual is that coordinated stream of personal experiences, which is the thread of life. It is that succession of self-realization, each occasion with its direct memory of its past and with its anticipation of the future. That claim to enduring self-identity is our self-assertion of personal identity (MT 161).

But Whitehead is well aware that this position – as far as it is developed at this point – is too simple to provide an acceptable solution to the problem of personal identity in humans: “Life is a bid for freedom: an enduring entity binds any one of its occasions to the line of its ancestry. The doctrine of the enduring soul with its permanent characteristics is exactly the irrelevant answer to the problem which

³ Of course the human stream of experiences, thoughts, emotions, [...] is far more complex than our remarks might suggest. But Whitehead’s account on what he calls ‘person’ or ‘soul’ leaves open the possibility of enrichment: “Our dominant inheritance from our immediately past occasion is broken into by innumerable inheritances through other avenues” (AI 189). So there may also be several more or less strongly distinguished serially ordered subordinate personal societies which might contribute to a reigning society. If we term the reigning society “I” and the dominated societies “selves”, we may even come close to Lakoff and Johnson’s (1999) position: “Lakoff and Johnson present convincing evidence that every metaphor for our inner life is a special case of a single general schema of ourselves as split. [...] In this nonconscious schema, a person is divided into an “I” and one or more *Selves*” (Galin 2003, 317).

⁴ For critical discussions and evaluations of Whitehead’s serial account of human persons see for instance: Johnson (1952), Edwards (1975), Fetz (1991, 2019); and for serial accounts of human persons in general: Langsam (2001), but also Hartshorne’s contrary position (1972b).

life presents. The problem is, How can there be originality?" (PR 104). So before we can go on, we must first deal with this obvious problem that seems to be inherent to Whitehead's position.

2.2.1 Identity vs. Novelty – a Dilemma

If societies – of course also including serially ordered personal societies – are defined as classes of actual entities which inherit a certain defining characteristic, this excludes the possibility of flexible adaptation to environmental changes, for this flexibility would lead to alterations and therefore to the destruction of the defining characteristic and hence to the destruction of the society in question. But on the other hand exactly this flexibility of novel reaction to environmental change is a major characteristic of *living* organisms and particularly of higher-order animals and human beings. In Whitehead's words, recourse to the defining characteristic in order to secure the identity of a society undermines the possibility of creative advance: "We ask for something original at the moment, and we are provided with a reason for limiting originality" (PR 104).

The problem can be stated as a valid logical argument: On the one hand, the identity of the human being has to be secured by conceiving of it as an enduring society. Identity is achieved if a defining characteristic is inherited by the next actual entity in the series and so sustained. But this solution – the monotone, rigid passing on of a defining characteristic – is exactly what becomes highly implausible if on the other hand the ability to flexibly manifest new characteristics in order to meet changing environmental influences constitutes a major characteristic of (at least) higher-order organisms, such as human beings. Actual entities which are members of a serially ordered society can only manifest the inherited defining characteristic and therefore are unable to act creatively, i.e. in a new way in response to environmental change. This paradox seems to have the logical form of a dilemma⁵: Either a human person (soul) is a serially ordered personal society or it is not a serially ordered personal society. If the human person (soul) is a serially ordered personal society, then the human person (soul) is *not* able to show flexible, i.e. *novel*, adaptive activities (since such activities would lead to a destruction of the defining characteristic). If the human person is not a serially ordered society, then the human person has *no iden-*

tity. So either the human person is *not* able to show flexible, i.e. *novel*, adaptive activities or it has *no identity*.

Of course neither of the two consequences is acceptable for Whitehead. On the one hand he maintains that an actual occasion functioning as a member of a personal society is indeed able to act adaptively: "[I]n each concrescent occasion its subjective aim originates novelty to match the novelty of the environment" (PR 102). In addition, he holds: "Yet personal unity is an inescapable fact. [...] Any philosophy must provide some doctrine of personal identity. In some sense there is a unity of life of each man, from birth to death" (AI 186).

So there seems to be only one way to escape this argument's conclusions: by questioning the truth of at least one of its premises. And, of course, Whitehead takes this direction. He introduces a distinction: He differentiates between serially organized personal societies on the one hand and what he calls a "*living person*" (PR 107 *italics added*) on the other. While any 'personally ordered society' has the form of a serial (chain- or thread-like) succession of single actual entities inheriting the defining characteristic from one member to the next in the row, a personal society is termed 'living' if it is able to adjust to changing environmental conditions by means of novel activities while still maintaining its identity within this welter of change. Whitehead terms these living personal societies "person" (AI 208), or "*living person*" (PR 107). The immediate surroundings of such a society – as already mentioned – are constituted by the brain. The brain represents an extremely complex environment of nested corpuscular and serially ordered societies, which are organized in such a way as to allow single members of the personal society to develop at least temporary conscious states.⁶ Consciousness in its different degrees opens the door for freedom: The iron grip of the immediate environment and its pressures is broken or at least loosened. Consciousness, which is based on the use of symbols and concepts, allows for expectations concerning future situations and hence for

⁵ For more details on 'constructive' and 'destructive' logical dilemmas' see: W. C. Salmon's *Logic* (1983, 68-71).

⁶ Consciousness according to Whitehead is a specific subjective form of an actual entity. More accurately it is a contrast between conceptual prehensions and physical prehensions; it is the felt contrast between that which might have been and that which in fact is the case. "Consciousness is the subjective form involved in feeling the contrast between 'theory' which may be erroneous and the fact which is 'given'" (PR 161). Whitehead also terms these complex prehensions 'intellectual feelings' (For more details see: Riffert 2003a & 2019).

planning how to (re-)act and adapt to future challenges. Whitehead terms such novel (re-)actions “conceptual initiative” (PR 102). While lower-order lifeforms are for the most part devoid of this capacity, “in the case of higher organisms, this conceptual initiative amounts to *thinking* about diverse experiences” (PR 102). According to Whitehead there may exist more than just one living society in one human being which may lead to a competition for influence among these societies. However, in healthy humans there is only one living personal society that dominates the hierarchy.

This living personal society – at least temporarily⁷ – provides (a) central direction to the bodily activities and (b) novelty of reaction to environmental ‘disturbances’.

It is by reason of the body, with its miracle of order, that the treasures of the past environment are poured into the living occasion. The final percipient route of occasions is perhaps some thread of happenings wandering in ‘empty’ space amid the interstices of the brain. [...] It receives from the past; it lives in the present. It is shaken by its intensities of private feeling, aversion or aversion. In its turn, this culmination of bodily life transmits itself as an element of novelty throughout the avenues of the body. Its *sole use* to the body is its vivid originality: it is the organ of novelty (PR 339 italics added).

So the advantage of a living personal society on top of the highly complex arrangement of the societies constituting the body consists in the opening of a hitherto unknown flexibility in reacting towards environmental changes. Or, “[t]he essence of life is the teleological introduction of novelty, with some conformation of objectives” (AI 206).

2.2.2 Strict Identity vs. Degrees of Identity

But the question still remains, and becomes ever more urgent: How is identity secured in such a flexible society? And more generally: What is meant by ‘identity’ in dealing with such a flexible society? In order to answer this question some more general

considerations regarding the term ‘identity’ are in order.

In the context of human persons, ‘identity’ cannot be conceived of in Leibniz’s sense of an ‘*identitas indiscernibilium*’, a concept which reaches back to Thomas Aquinas, Plotinus and even Aristotle, and which was later introduced into the domain of modern formal logic by Frege and Russell (see: Burkhard 1984, 89). Leibniz’s strong concept of identity implies an ‘absolute equivalence of predicates’ since two objects are indiscernible if and only if all the predicates which describe the attributes of one of them can also be used without any qualification to describe the other object and vice versa (without any change concerning the truth values of the statements in question). This characterization of identity makes it transitive: if A is identical with B and B is identical with C, then A is identical with C. It further means that we can without any restrictions substitute one of the two identical ‘objects’ for the other. Identity in this sense is an ‘all or nothing’ affair. One consequence of this, of course, is that it excludes the possibility of change, for if a new predicate were ascribed to a ‘thing’ or an ‘object’ then a new identity would have been created, as Hartshorne noticed. He also drew attention to the consequence of such a conception of identity: “Leibniz on the other hand explains endurance too well. He virtually denies change altogether. [...] For if nothing can enter the substance which is not already a part of its identity, then to explain its adventures no other substance need or can be considered” (Hartshorne 1972a, 50).

Whitehead is critical about this strong or pure sense of ‘identity’. His critical comments on pure identity can even be traced back to his early writings in the field of mathematics, a field where the strict meaning of ‘identity’ finds its strongest support.⁸ Whitehead here holds the position that pure identity at best is the *ideal border case* of equivalence (UA 5f); a kind of ‘absolute’ equivalence while ‘partial’ equivalence is to be conceived of as implying only a partial identity. So Whitehead in a sense draws a distinction between equivalence and identity.

In his book, *A Treatise on Universal Algebra*, he uses the sign ‘=’ to indicate an equivalence between the groups of signs positioned on both sides of it: “One thing or fact, which may be complex and involve an interrelated group of things or a succes-

⁷ During periods of sleep, powerlessness, unconsciousness or coma/persistent vegetative states, the autonomic nervous sub-systems execute life-securing functions without the active involvement of the ‘reigning society’, which dominates the hierarchy during periods of consciousness (see for instance: MT 25, PR 108).

⁸ In dealing with this topic I have strongly profited from Wolfe Mays’ book *Whitehead’s Philosophy of Science and Metaphysics* (chapter 3: ‘Equivalence’, 19-32).

sion of facts, is asserted to be equivalent *in some sense or other* to another thing or fact" (UA 5 italics added).⁹ The phrase 'equivalent in some sense or other' also makes clear that equivalence for Whitehead necessarily is context specific: "[t]he equivalence of distinct things implies *a certain defined purpose in view*, a certain *limitation* of thought or action" (UA 5 italics added). The use of the term 'equivalence' hence presupposes a certain context in which it is relevant. In this sense it is context-relative.¹⁰ If we express that two propositions are equivalent, ' $x = y$ ', the context relative to which this assertion is made must also be explicitly stated. Whitehead is very clear about this also in his book, *The Axioms of Projective Geometry*. He maintains that the statement 'The angles at the base of an isosceles triangle are equal' is deficient. It has to be supplemented by a phrase that specifies the context in which it is supposed to hold: "Such and such axioms imply that the angles at the base of an isosceles triangle are equal" (APG 2). This idea of context-specificity is still present in Whitehead's later philosophical work: While two persons may have the same measured IQ, they may yet differ in respect to their religious beliefs or the colour of their hair. This also holds for one person at different times: Person x at time t_1 (6 months old) may be equivalent to the person x at time t_2 (75 years old) *in respect to* the place and time of birth or *in respect to* his/her parents. In *other respects* they may totally differ – for instance in the ability to solve an equation of the second degree.

In line with this position, Whitehead distinguishes between the sign '=' and the copula 'is'. While the former indicates an equivalence (or partial identity), the latter maintains pure identity (or absolute equivalence). In pure mathematics or logic, absolute identity, as simply introduced by arbitrary definitions, may not be a great problem (although this can be disputed). It certainly is when we are concerned with applied mathematics and applied logic. Indeed, following Bradley (1922, 374), if we ascribe absolute identity to two different things we first have to distinguish them (at least in respect to space and/or time); and so they are (by definition) not identical in Leibnitz' absolute or pure sense of

an 'identitas indiscernibilium' (because they differ at least in respect to space and/or time). This seems to be the reason why for Whitehead there can be no absolute identity between two actual entities: Since in his metaphysical approach, each of them defines a different locus in space and time, and thereby different perspectives on the world, this differentiates them from one another (not only in space and/or time, but also qualitatively in terms of the way they perceive the world, and therefore in terms of the the eternal object they realize).

Whitehead consequently remains critical in his later works concerning the applicability of such a logic (i.e. a logic in which this sort of absolute identity is presupposed) to real world phenomena. The term 'variable' in traditional formal systems, according to Whitehead, rests on this kind of concept of pure identity: "Also the variable, though undetermined, sustains its identity throughout the arguments. [...] The use of the variable is to indicate the self-identity of some use of 'any' throughout the train of reasoning" (MT 106f). This implies, according to Whitehead, two tacit presuppositions which are both questionable: "[O]ne is that the definite symbols of composition can retain the same meaning [i.e. strict identity] as the reasoning elaborates novel compositions. The other presupposition is that this self-identity of each variable can be preserved when the variable is replaced by some definite instance" (MT 107).

From the perspective of Whitehead's radical process approach both presuppositions are called into question. Whitehead draws a skeptical conclusion concerning the possibility of absolute identity: "Complete self-identity can never be preserved in any advance to novelty" (MT 107).¹¹ He does so

⁹ Of course Whitehead here uses the term 'thing' in a very broad manner; it covers any objects such as "concrete or abstract, material things or merely ideas of relations between other things" (UA 19).

¹⁰ This concept of context-specificity is also present in Whitehead's ethical ideas, which puts him close to ethical situationism (see: Riffert 2003b, 180).

¹¹ Rescher in dealing with the same problem proposes the development of a temporal logic as a tenable solution: "The static *being* of the Western theory of logic and language has led various philosophers ... to deny the adequacy of these instrumentalities vis-à-vis a world of activity and change. But rather than simply deny the competency of logic and language per se, it would seem more plausible to argue for their extension and enhancement. And this [is] the direction in which modern developments have, in fact, moved. Some account of these developments is given in Nicholas Rescher and Alistair Urquhart, *Temporal Logic* [1971]" (Rescher 1996, 193-194 fn.). Whitehead too had expressed an unbroken optimism concerning the potentials of logic and mathematics: "Mathematics is the most powerful technique for the understanding of pattern, and for the analysis of the relationships of patterns. ... Having regard to the immensity of its subject matter mathematics, even modern mathematics, is a sci-

although he is very well aware that this implies a dangerous, even disastrous consequence:

A difficult problem arises from this doctrine. How can the notion of any generality of reasoning be justified? For if the process depends on the individuals, then with different individuals the form of process differs. Accordingly what has been said about one process cannot be said of another process. The same difficulty applies to the notion of identity of an individual conceived as involved in different processes. *Our doctrine seems to destroy the very basis of rationality* (MT 97 italics added).¹²

Whitehead draws the conclusion that within his metaphysical system the absolute general applicability of logic and mathematics (and in fact of any formal system), which is often seen as providing a secure basis for induction, is destroyed (MT 98). In trying to save rationality he retreats to analogy. Although it cannot be denied that inescapable diversities are the limitations of rationality, he claims that all these obvious “diversities are not absolute. Analogies survive amid diversity” (MT 98).

This of course also makes it necessary for Whitehead to relativize the role of the application of the formal sciences in the process of metaphysical research: “We thus dismiss deductive logic as a major instrument for metaphysical discussion. Such discussion is concerned with the eliciting of self-evidence. Apart from such self-evidence, deduction fails. Thus logic presupposes metaphysics” (MT 107). According to Whitehead, applied formal sciences, such as logic and mathematics, have to be introduced on a pragmatic basis: “When applied to concrete instances, it is a tentative procedure, finally to be judged by the self-evidence of its issue.” (MT 106).¹³ So whether a logical argument can be applied

meaningfully in and to a specific context is a matter of degree: “The only question is, as to whether the loss is relevant to the purpose of the argument” (MT 107). Here again Whitehead introduces the idea of ‘context-specificity of equivalence (partial identity)’, which was already present in his mathematical works: It is important to specify whether the loss is relevant in the context of our reasoning! In the end (after having considered many other important contextual constraints), this judgement is relative to a certain metaphysical standpoint which defines what is relevant within a certain context and what is not.

The strong or pure concept of identity, according to Whitehead, can neither be applied in everyday practice nor in mathematics nor in any other scientific or philosophic research field. Here Whitehead seems to side with Rudolf Carnap, one of the great figures of Neo-positivist thinking, who also had noticed this fact: “Identity in everyday language usage and also in the sciences, is not taken in its strict meaning. Also things, which in the strict logical sense are not identical, are *linguistically treated as identical*” (1961, 217 my translation, italics added).

This implies that when dealing with the question of diachronic identity we are dealing with partial identity. Of course, the man who committed a crime *decades ago* and who is on trial for his deed *now* has changed during these years. The man today by no means is in the strict Leibnizian sense identical with the perpetrator many years ago. And the same holds for ourselves today and ourselves as babies some decades earlier. According to Whitehead there only holds partial identity (equivalence) between the man *x* at time *t*₁ and the man *x* at time *t*₂. In respect to certain features they may be partially identical, while in respect to other features they may not. Identity in this case is a matter of degree, and context specificity is essential to judgements concerning identity.

“The word ‘*continuing*’ states only half the truth about human identity. In one sense it is too weak, and in another sense it overstates. It is too weak insofar as we not only continue,

ence in its babyhood. If civilisation continues to advance, in the next two thousand years the overwhelming novelty in human thought will be the dominance of mathematical understanding” (ESP 109).

¹² This problem as we shall see in (3.) also has implications in the domain of ethics.

¹³ But all this does not imply that the formal sciences are unimportant and can or should be neglected. What Whitehead stresses is the fact that the formal sciences do not lead to *certainty* regarding truth in research. Nevertheless, they are still important auxiliary instruments. Their value lies in the fact that they expand the realm of testing in that they allow for the deduction of consequences (from the axiomatic matrix) which are empirically (external contradiction) and logically (internal contradiction) testable.

They extend the possibilities for a theory to be tested by ‘self-evidence’. For a more detailed exposition of the interplay between the formal sciences, factual sciences and metaphysical systems from Whitehead’s perspective see: Eastman 1997, 1998, 2004; Eastman & Keeton 2004; Eastman & Riffert 2009; Riffert 2004a, 2004b; Riffert & Cobb 2003; Riffert & Eastman 2008; Riffert & Sander 2008; Riffert & Weber 2003.

but we claim absolute identity with our previous state(s). We identify wholly with the self that was (formerly) in a particular state of mind, and it is this state of mind that serves as the basis of our present experience (e.g., a quarter of a second later). But in another sense the word '*continuing*' is too strong. Our experience and states of mind are marked by discontinuity and change. New elements have intervened. All of these new elements are provided by our bodily functions and interactions with our environment. We fuse these new elements with the basic stuff of experience provided by our previous states of mind (e.g., our state of mind a quarter of a second ago) and thus our present state is constantly changing (MT 160)."

Before we further discuss this notion of equivalence or partial identity in the context of human diachronic personal identity, we will first turn to the question of whether there is some empirical support for the claim that human identity can degenerate or even get lost.

2.3 Fragility and Loss of Identity

Identity, even partial self-identity, is not pregiven in Whitehead's approach but has to be actively constructed by fusing antecedent conditions with the actual conditions of our life. So, my partial self-identity emerges as the synthesis of my immediate past as a 'living personal society' and of the new stimuli imposed on me by the brain mediating the impacts of a more distant environment. If a synthesis is reached that shows some continuity of pattern with the immediately passed actual entity of the 'living personal society', then identity is partially realized and secured. But such synthesising aimed at conserving patterned continuity can fail; even more so since this act of construction, which in Whitehead's terms is the partly self-creative concrescence of an actual entity, depends strongly on its favourably structured environment, especially the brain.

The inflowing stimuli, provided by the more distant surrounding and mediated as well as transformed by the brain, may be such that it will not be possible to 'fuse' these new elements into a coherent and continuing unity with the preceding actual entity of the sequence of actual entities which have built the 'living personal society' up to this moment. This painful truth is demonstrated strikingly in some neurological and psychological pathologies.

The far reaching dissociations, so-called 'de-realization' and 'de-personalization' (Süllwold 1983), during acute episodes of schizophrenia or a condition called 'multiple personality disorder' or 'dissociative identity disorder' – the latter being well-documented in recent clinical psychiatric research (for an overview see: Putnam 1991, Brand et al. 2016) – are frightening, yet strikingly apt examples of this phenomenon. A theory that has increasingly garnered support – although it remains only one among several alternative and even competing hypotheses – is that this failure may be due to biochemical changes in the brain related to the production or distribution of neurotransmitters like dopamine (McCutcheon et al. 2020). Disturbances in dopamine production may cause a massive 'flooding' of incoming stimuli which is impossible to integrate into a unity (Süllwold 1983). Thus, it seems to be the case that schizophrenia is caused by an imbalance in neurotransmitters that may lead to an uncontrollable rush of incoming stimuli which cannot be integrated into the concrescence of an actual entity. Such uncontrollable 'overflow' of input may either lead to a break down issuing into catatonic stupor or to a splitting off of some of this incoming information into non-personal, external and 'foreign' states (for instance, one person may not be able to construct him-/herself as the owner of his/her own thoughts). In certain cases, one can speculate that this may even lead to a branching of the serially ordered society into different strands, as in the case of 'multiple personalities'. This is what one may expect from a Whiteheadian point of view: "There are limits to such unified control, which indicate dissociation of personality, multiple personalities in successive alternations, and even multiple personalities in joint possession" (PR 107).

In other, less severe, neurotic cases, malfunctioning cognitive processing may lead to a rigid blocking of new and unfamiliar input because of fear and (unrealistically) expected danger to one's own life or well-being. In these latter cases, patients have the impression of needing to secure diachronic identity by blocking out endangering input. Here, partial diachronic identity or stability is secured at the cost of adequate novel, adaptive reactions to the environment. Whitehead's philosophical system can account for such phenomena given his concept of negative prehension: "A negative prehension holds its datum as inoperative in the progressive concrescence of prehensions constituting the unity of the subject" (PR 23-24). Such blocking of inflowing information leads – at least – to the impoverishment

of the individual, and in the long run, may even endanger their diachronic identity by leading to the development of major depression, which in some extreme cases may even end in suicide.

Further important empirical support for the possibility of loss of identity comes from the so-called ‘split-brain experiments’ undertaken by Roger Sperry and Michael Gazzaniga among others. (For good accounts see for instance: MacKay 1987; Galin 2003.) Parfit too (1986, 1987) discusses the case of a split-brain patient who, after undergoing an operation that resulted in the severing of his corpus callosum (the ‘data-highway’ between the two brain hemispheres), exhibited cognitive changes that appeared to undermine the unity of personal identity. The patient received in the right hemisphere of the brain from the left visual field via the right nervi optici of both eyes the stimulus of a red patch, and in the left hemisphere of the brain received from the right visual field via the left nervi optici of both eyes the stimulus of a blue patch. If asked to write down with each hand separately how many colours he saw, the patient (with each hand) wrote down: ‘Only one colour’. If further asked: ‘Which colour do you see?’, he wrote with the right arm (which is controlled by the left hemisphere): ‘Red.’ And with the left arm (which is controlled by the right hemisphere): ‘Blue’. This suggests the presence of two different perceptions at the same time. Therefore, Parfit claims that “the division of one stream of consciousness into separate streams, has [...] happened” (1987, par. 31). This position gains further support from the fact that split-brain patients report that the right and the left hemispheres seem to make different decisions issuing in two different actions. For example, a split-brain woman reported that when she had gotten up one morning she (consciously) decided to wear a green dress; but when her right hand (controlled by her dominant and conscious left hemisphere) reached for the green dress, simultaneously her left hand (controlled by her unconscious right hemisphere) reached for the red dress. This indicates that in split-brain patients there are not only two different streams of perception, but also two different instances of decision making, suggesting a divided *locus of control*, which could itself be termed person or self. Such phenomena are hardly explainable within static substantialist accounts of personhood and strict identity.

Further examples are severe cases of retrograde amnesia and dementia in general. Severe retrograde amnestics may almost completely (at least con-

sciously)¹⁴ be cut off from their former life, so that they cannot even remember their spouse, children and parents and can manage their everyday life only with external help. In these cases, dramatic damage to the brain due, for instance, to some accident or sickness, has caused a severe mnemonic disruption of personal continuity.

We can sum up this short excursion into the empirical state of affairs concerning the endangerment of human diachronic identity with Whitehead’s words: “The world is thus faced by the paradox that, at least in the higher actualities, it craves for novelty and yet is haunted by terror at the loss of the past, with its familiarities and its loved ones” (PR 340).

Without going into further empirical details it should have become clear that Whitehead’s position (a variable partial diachronic identity), which implies the possibility of improvement but also the possibility of degeneration or even loss of (partial) self-identity, rests on and is supported by a great variety of empirical data from scientific experimentation and controlled clinical observation.

2.4 Sources of Partial Identity

Now that it has become evident that human diachronic identity cannot be understood in the strict (Leibnizian) sense but rather in a sense that is only partial, fragile and subject to dissolution, we shall turn to the even more pressing question of what it is that secures at least the partial diachronic identity of a person. According to Whitehead there are two main sources of our self-identity: (a) the body and (b) the living personal society itself – the person, of which our immediate moment is its latest member.

Thus our experience in the present discloses its own nature as with two sources of derivation, namely, the body and the antecedent experiential functionings [of our living personal society]. Also there is a claim for identification with each of these sources. The body is mine, and the antecedent experience is mine. Still more there is only one ego, to claim the body and to claim the stream of experience. I submit that we have here the fundamental

¹⁴ There may, however, remain some implicit memory as research during the last few decades has substantiated. Perrig & Wippich (1993) provide an extensive overview. Also of special interest here are Weiskrantz’s research results (1997) on the so-called phenomenon of ‘blind sight’.

basic persuasion on which we found the practice of our existence (MT 160f).

Let us first turn towards the body (including, of course, the brain). According to Whitehead, the human person (living personal society) as a continuously enduring entity is dependent on its body (including the brain). "All sense perception is merely an outcome of the dependence of our experience upon bodily functionings" (MT 159). Since there is no clear-cut boundary between the body and the environment it occupies, to a certain albeit lesser extent, a person's social and natural surroundings also support their existence and their continuity over time. In turn, the streams of perceptions, volitions, thoughts, etc. of the regnant personal society provide a unified direction to the body. Thus, persons are very intimately connected to their bodies in a reciprocal way. Consequently, any central malfunctioning of the body (brain), as mentioned in the last section, endangers the (partial) self-identity of the human person. In this sense the body is a *conditio sine qua non* of human identity. It may play an important role in securing partial identity, at least as long as it is functioning well. However, if the body's functions are impaired, our identity may degrade or even get lost.

But, as Whitehead claimed, there is still another source of identity in humans: the living personal society itself. While our brain and body (and even nature as a whole) present a *conditio sine qua non* of partial personal identity in that they establish more or less stable regularities, the serially ordered living personal society has to *construct* the specific (partial) identity by integrating novel stimuli with its own inherited defining characteristic. In mankind, the dominant dependence on bodily feeling seems still there. Yet the life of a human being receives its worth, its importance, from the way in which unrealized ideals shape its purposes and tinge its actions (MT 27).

John Locke held that a person's explicit or conscious memory alone could secure identity: "[A]s far as this consciousness can be extended backwards to any past Action or Thought, so far reaches the Identity of that Person" (Essay, II, XXVII, § 9). His approach became influential up to our present times, even though Joseph Butler (see for instance: 1975 [1736]) and Thomas Reid (see for instance: 1975 [1785]) had already criticized this account by pointing out that we can, for instance, forget our earlier

experiences as well as our thoughts and actions.¹⁵ This implies that when we lose the conscious grasp of our memories – for instance, while sound asleep, in a coma, during acute episodes of schizophrenia, or as a result of brain degeneration, e.g. in Alzheimer patients – our identity will fade along with our ability to consciously memorize.

In order to meet this critique, the concept of memory has to be extended: The concept of explicit, conscious memory has to be supplemented by what is called in contemporary psychology nonconscious or implicit memory. The contents of implicit memory can be conceived as nonconscious i.e., non-activated, memory traces, which were caused by earlier perceptions, thoughts or actions. The contents of explicit, conscious memory on the other hand are temporarily activated memory traces. At least during the last decades scientific research on implicit memory seems to have substantiated the claim in favour of such an extended concept of memory. (For an early and a current overview see: Schacter 1987; Mulligan 2008; and for an application of this concept of implicit memory in the field of dealing with patients suffering dementia see for instance: Fuchs 2020.)

Whitehead's concept of the 'human person' as a thread of successive actual entities very naturally accounts for such an extended concept of memory securing physical identity, since it includes an immediate and unconscious transference of information from actual entity to actual entity by causal efficacy (i.e. implicit memory). "Memory is an example of perception in this mode [of causal efficacy]. For memory is perception relating to the data from some historic route of ultimate percipient subjects M1, M2, M3 etc., leading up to M which is the memorizing percipient" (PR 120). Only from time to time may some memorized information reach the level of consciousness (i.e. explicit memory), explicitly necessitating novel reactions to new intruding stimuli. So basically, "causation and *physical memory* spring from the same root: both of them are physical perception" (PR 239 *italics added*).

But what differentiates the causal inheritance of the members of a living personal society, a human

¹⁵ It is interesting enough to note that even the eminent process philosopher and theologian, John Cobb (2007, 35-40) accepted John Locke's position: "My sense of personal identity seems to me to depend upon [conscious] memory. ... Only [conscious] memory can serve in my self-understanding to determine self-identity through time" (2007, 36). For a critique of Cobb's position see: Bennet (1973).

person, from all the other inheritances taking place among all the billions of actual entities which are not members of a particular living personal society? We still have not dealt with the question of what it is that secures at least partial temporal identity since the concept of continued (unbroken) connection *per se* does not yet account for a *qualitatively distinct continuity* between the serially connected occasions of a particular living personal society, which at the same time are able to flexibly react to new stimuli flowing in from the environment.

Whitehead nowhere in his oeuvre gave a detailed account of partial human diachronic identity as based on the living personal society itself. But one can find two hints in Whitehead's work which at least provide a glimpse of Whitehead's proposed solution: steadiness of purpose and what he had termed the "peculiarly full summation of its [the actual human person's] predecessors" (S 27).

Let us first turn to the steadiness of purposes: There is one passage in which Whitehead writes: "The essence of life is the teleological introduction of novelty, with some conformation of objectives. Thus novelty of circumstance is met with novelty of functioning adapted to *steadiness of purpose*" (AI 206 italics added). While humans respond to the novelty of environmental stimuli by generating novel reactions, they do so in order to secure the steadiness of purposes. So it seems that Whitehead holds that steadiness of purpose contributes to the partial identity of humans over time.

According to Whitehead there are purposes or (subjective) aims active in every actual entity, down to the most primitive forms. But these aims or goals in the primitive forms of actual entities usually are only blindly received from past actual entities and/or from God. These aims are bound to the ongoing rhythm of nature and therefore only passively contribute to the repetition of pre-existing patterns. Only at the conscious level, which is triggered when unconscious automatic adaptation to the surroundings fails, do these aims gain importance for dealing with novelty. They can be projected into the more or less distant future and so can provide steadiness and continuity to the organisms' advance into their own future realizations. So, in the highly developed human organism what continues as identical may be the (more or less) freely chosen purposes or aims. Humans, by ascending to the level of conscious thought, gain the (although still limited) ability to determine their future by introducing persisting purposes amidst the continuous changes of reality.

Whitehead discussed one domain of human life in which purposes are central in some more detail: ethics. It seems to be a characteristic of human beings to organize their enduring lives around stable ethical values. Such values, if accepted by a specific culture, may be codified, as in fact they are in the Ten Commandments of Judaism and Christianity or the Noble Eightfold Path of Buddhism. Both of them, for instance, strictly prohibit (at least) the killing of human beings, and the followers of these two religions try to organize their lives around this fundamental aim.

However, environmental pressures may be extremely strong, and new intruding elements (stimuli) may hardly be compatible with the steady personal goals of an individual. In such cases, it may even be necessary to give up not only peripheral aspects but also more central parts of the person's core aims in order to preserve at least some other core aims in the face of invading new information. What this core characteristic may be will differ largely from person to person; to determine the purposes of a particular human being and their place within his/her value system is an empirical question. The core purposes of a specific person may consist of religious beliefs or more fundamental concerns; for instance, in the dehumanizing context of a concentration camp, the individual's core purposes may be reduced to the absolute minimum of pure physical survival. Under such extreme circumstances it may be necessary to give up many if not all other values in order to preserve the most important one. Some may decide that physical survival is the most important value, while others may continue to prioritize their moral and/or religious convictions – for example, Father Maximilian Maria Kolbe (OFMConv), who sacrificed his life to save that of another, a husband and father of two children in the concentration camp of Auschwitz in August 1941. Thus, our convictions – even our central purpose of life – may or may not change due to influences from our surroundings.

Whitehead was well aware of this fact, which is also the reason why moral codes do not hold without exceptions. For instance, in relation to the Ten Commandments, he wrote:

[T]hey are formulations of behaviors which in ordinary circumstances, apart from very special reasons, it is better to adopt. [...] The *destruction of a man*, or of an insect, or of a tree, or of the Parthenon, *may be moral or immoral*. The Ten Commandments tell us that in the vast majority of cases such slaughter is better

avoided. In these exceptional instances we avoid the term 'murder'. Whether we destroy, or whether we preserve, *our action is moral if we have thereby safe-guarded the importance of experience so far as it depends on that concrete instance in the world's history* (MT, 20 italics added).

A related discussion of human values can be found in Whitehead's *Science and the Modern World* in which he emphatically advocated the importance of intercultural exchange for the development of mankind: "A diversification among human communities is essential for the provision of the incentive and material for the Odyssey of the human spirit. Other nations of different habits are not enemies: they are godsend" (SMW 207). As important as this intercultural exchange may be, Whitehead again is well aware that there are limits to intercultural exchange, since it may, in extreme, exceptional cases, undermine the possibility of the advancement of humanity: "It must be admitted that there is a degree of instability which is inconsistent with civilisation" (SMW 207). Indeed, the stability provided by civilisations is indispensable for creating complex, nuanced aesthetic experiences.

These two examples clearly show how close Whitehead's ethics comes to situationist ethical approaches (for instance, the approach advocated by John Dewey (1948)). In short, in a world that is characterized by continued change, decisions have to be made *in concreto*, reflecting both the relevant moral codes as well as the specific conditions of the actual situation in question.

All this does not mean that steadily held moral codes have no important function. First, they indicate what behavior under ordinary circumstances leads to an intensification of aesthetic feeling for the vast majority (or indicate what prevents the decay of intense aesthetic experiences for the vast majority). Moral codes thereby release us from the burden of reasoning and evaluating in each and every new situation in order to determine which behavior might morally be best. So, in the domain of ethics, Whitehead's well-known dictum, namely, : "[c]ivilization advances by extending the number of important operations which we can perform without thinking about them", is correct (IM 46). Second, moral codes lead to the stabilization of social life, which is the *conditio sine qua non* for aesthetic advance: "Morality consists in the control of process so as to maximize importance. It is the aim at greatness of experience in the various dimensions belonging to

it" (MT 13f). Since, according to Whitehead's process metaphysics, striving for aesthetic enjoyment is always possible and to be aimed at, violations of moral commands and prohibitions are not only possible but – under certain (extreme) conditions – even demanded if it is the case that they "will promote the evolution of that environment towards its proper perfection" (AI 292). Moreover, since this evolution towards the intensification of aesthetic feelings is an unending process, there is no space in Whitehead's metaphysics for the concept of a 'highest good', a *summum bonum*: "There are in fact higher and lower perfections ... [However, a]ll realization is finite, and there is no perfection which is the infinitude of all perfections" (AI 257).

In addition to external pressures, there may also be internal pressures which force a person to rearrange his/her hierarchy of normally steady convictions and purposes. We all are aware that we hold different goals. Some of them may be more or less central to the core of a certain person's value system, and some of them will be instrumental to the realization of a few higher or more central aims. These more central or higher goals, which also implicitly embody and express a person's core values, build what I would like to call the 'qualitative core' of each person's identity. Of course, there is every reason to suppose that these goals are person-specific. This implies that there is no possibility of providing a general account of the content of the qualitative core that could be applicable to each and every person. Such an account, however, is not necessary since psychologists have developed and still are developing diagnostic tools which at least partially allow for the detection of idiographically organized hierarchies of goals. (See for instance: George Kelly's Rep-Grid technique (1955) or – more elaborated – Franz Caspar's 'plan analysis' (1995 & 2007).)

Rearrangements of personal (core) goals may be necessary for the sake of the psychic and physical well-being or even survival of a person. Thus, even core convictions can become a topic of change in the psychotherapeutic process in order to promote growth and health in the individual. In such therapeutic processes it may be necessary to advance the client's competences for flexibly adapting to the changing environment in order to successfully maintain the steadiness of their purposes, or – if this is not possible (because of a dilemma between two of his/her goals) – to reformulate their purposes, even those which are most important or central to their

identity. In either case, contradictions between purposes have to be transformed into contrasts.

So again: Even though the steadiness of (central) purposes may contribute to the partial identity of human beings, goals and purposes also change over time and sometimes have to change in order to secure the person's survival or reality's advance into novelty of enhanced aesthetic enjoyment. So we are left with a final hint which Whitehead gives concerning the establishment of the partial identity of human beings.

Whitehead's hint is presented in the following three quotes: "We – as enduring objects with personal order – objectify the occasion of our own past *with peculiar completeness* in our immediate present" (PR 161, *italics mine*). Similarly: "An enduring personality in the temporal world is a route of occasions in which the successors *with some peculiar completeness sum up* their predecessors" (PR 350, *italics added*). Again, but now explicitly connecting 'peculiarly full summation' with identity: "These concrete moments [of a living personal society] are bound together into one society of *partial identity* of form, and by the *peculiarly full summation of its predecessors* which each moment of the life-history gathers into itself" (S 27, *italics added*).

All three passages seem to suggest that, in Whitehead's view, there exists an alternative to the continuous and rigid inheritance of a defining characteristic in non-living personal societies in order to secure partial identity: It is the origination of a new pattern, which does not exclude but rather includes the old defining characteristic *as well as* the new environmental elements. The continuity between the older defining characteristic and the new one is thus to be characterized as an improved patterned contrast. The new actual entity forms a new pattern in which the old defining characteristic is objectified 'with a peculiar completeness' along with the new elements. The old defining characteristic is extended in order to include the new elements; this is achieved by differentiating the old defining characteristic. In a way, this may be similar to an accumulative scientific progress: A new theory (ideally) should include and accommodate the merits of its predecessor but should also be able to explain new facts. Whitehead holds a similar position in respect to progress in cosmology: "At certain epochs a cosmology may be produced which *includes its predecessor* and assigns to them their scope of validity" (FR 88 *italics added*). Still, in regard to either cosmology or personal development, this inclusion will not usually result in perfect completion. New cosmologies will generally

"[fail] to include some of its [the old cosmology's] general truths" (FR 88). And – as we have seen – the same is true for the human individual: The newly developed pattern may not embrace all aspects of the old pattern, nor all new elements. Whitehead seems to refer to this fact by the phrase, "with *some peculiar completeness*" (PR 350).

That under normal conditions we are aiming at differentiating our already existing patterns in order to sustain them and yet be able to integrate new elements has been shown by Jean Piaget. Piaget calls this tendency '*l'équilibration majorante*' [heightening equilibration] (Piaget 1976b, §6, 36-44). Similarly to Whitehead, he conceives the world as consisting of dynamic self-regulatory structures. He emphasises the processual character of the structures, their constant aiming at new equilibrium, which he terms equilibration or structurization. For him the human organism is a complex arrangement of structures, or rather structuralizations. On top of this complex of structures which are hierarchically ordered, there is a linear and unbroken sequence of structuralizations, which partially dominates the organism as a whole and provides it with a unified directive purpose.¹⁶ The tendency towards the establishment of an equilibrium, however, is present at all levels of reality – from the sub-atomic structures to the mental concepts of human beings. As an example, the development of a particular sensorimotor schema of a young child may suffice here for demonstration: A child at the age of about 6 months has already developed a behavioral schema (or sensorimotor structure) that allows for the grasping of a rattle. This child can without any problem apply this schema to things similar to the rattle. However, if the child tries to grasp very tiny objects – for instance a ring – the schema has to be modified in order to adapt to the new features of tiny objects; the child now has to use the fingertips of the thumb and index finger to be able to grasp the small object. A new structure (with a new equilibrium) has now developed. This newly developed behavior structure now includes the possibility of grasping rattle-like objects as well as smaller things. So, in the new structure, the older abilities have been preserved and

¹⁶ For a detailed comparison of the striking parallels between the basic concepts of Whitehead's process philosophy and Piaget's general genetic structuralism, see Riffert (1995, 2003c & 2007b, 2009). A creative attempt to fuse Whitehead's process philosophy and Piaget's genetic structuralism can be found in Reto Fetz's recent book *Die Wirklichkeit der Wirkwesen* [The Reality of Efficient Entities] (2019).

are supplemented by new ones. An *équibration majorante* has taken place. The newly developed motor structure of 'grasping' is partially identical with its predecessor in that it covers *all* the potential behaviors of the former. But, in addition to this identity with the older structure new abilities are gained and included in the new structure as well. This is made possible by a qualitative change (or improvement) of the older structure, which leads to an enrichment without any loss (or very little loss). This encompasses a kind of *accumulation* of abilities during the development of the child's structures. If we now apply this to the (partially) dominating sequence of structures we can account for identity. Although this identity is only partial, the identity is secured by summing up the abilities inherent in the predecessor (older structure). Piaget conceives the whole cognitive development of the child as a sequence of such improving equilibrations. Consequently, he refers to this *équibration majorante* in order to account for the (partial) identity of an organism in general: "[T]herefore assimilation is the factor of permanence and continuity of the organism's forms" (Piaget 1968, 70 author's translation).¹⁷

Piaget's stage theory of the child's cognitive development rests on a wide range of empirical material. His stage theory contains the following assumptions: (1) there are different stages of (sensorimotor and cognitive) development, (2) the stages have a definite order which cannot be reversed (though different individuals may need different amounts of time to pass through each stage), (3) the stages are ordered from less complex to more complex, (4) each stage builds upon the previous stage, differentiating and extending it, and (5) therefore, no stage can be skipped in development but is a necessary condition for reaching the next stage. Similar stage theories (which all rest more or less explicitly on the idea of the *équibration majorante*) were proposed by others, for instance by Lawrence Kohlberg (1981, 1984) concerning the development of moral judgement, and by Fowler (1981) as well as Oser and Gmünder (1988) in respect to the development of religious judgement.

¹⁷ It is interesting to note that Piaget's *équibration majorante* and Whitehead's 'summation with peculiar completeness' seem to echo Hegel's central concept of 'Aufhebung', a lifting up of previous knowledge into an externally more embracing and internally more differentiated – and therefore more complete – higher unity of the developing 'Weltgeist' [world spirit]. In Whitehead's metaphysics, God's consequent nature accounts for this function of Hegel's *Weltgeist*.

Still, Piaget and Whitehead are both well aware that this process of an *équibration majorante* may fail. In this case, loss of abilities – even decay which may endanger the identity of the organism – may result, for example, in the case of late-stage Alzheimer's disease.

So both ideas – (a) steadiness of purpose and (b) summing up the immediate past with peculiar completeness – contribute to the creation of the partial identity of a human person in Whitehead's sense of a living personally ordered society. But it is important to bear in mind: it is only partial identity at best, and even partial identity is still constantly threatened since integrating new incoming stimuli into a new unity may necessitate the negative prehension of formerly positively prehended contents. Hence, some previously available contents will get lost. Here is no final security but only adventure, or as Whitehead puts it: "The world is thus faced by the paradox that, at least in the higher actualities, it craves for novelty and yet is haunted by terror at the loss of the past, with its familiarities and its loved ones" (PR 340).

3. Final Remarks

To conclude, according to Whitehead, what secures the continuity of the human person are: (a) the overlapping of actual entities from the human body and brain (as a necessary but not sufficient condition), (b) the continuance of a pattern of core purposes or goals, and (c) the tendency in organisms towards an *équibration majorante*. None of these conditions seems to be able to *entirely* secure the diachronic partial identity of the human being. Life remains adventurous – even more so in our modern times: "Modern science has imposed on humanity the necessity for wandering. Its progressive thought and its progressive technology make the transition through time, from generation to generation, a true adventure" (SMW 207).

The extent to which continuity in a concrete individual person is sustained can only be estimated empirically and idiographically: either directly by the person in question, i.e. from the first-person perspective, or indirectly from the 'outside', i.e. from the third-person perspective – either through a naïve inference based on overt behavior or with the help of sensitive diagnostic tools developed for idiographic measurement of a person's idiosyncratic cognitive features and value systems. This conclusion, for some, may prove unsatisfying, but it is – as far as I can see – plausible in the face of empirical facts –

those provided by scientific research (in psychology and neurophysiology) and by clinical case studies (in medicine and especially psychiatry), but also those furnished by observation of and participation in everyday experience. Human persons, as constructed within Whitehead's approach, seem to be characterized best as 'Hobos': They are homeless boys and girls searching for sheltering stability amidst the uncertainty of an environment in constant flux.

However, Whitehead's organismic process philosophy offers a faint hope for the Hobos: On the one hand, it recognizes and emphasizes the soul's dependence on the functions of body: "[E]very type of crucial experiment proves that what we see, and where we see it, depends entirely upon the physiological functioning of our body" (MT 159 italics added). On the other hand, it sketches a possibility for how this dependence could be overcome and, thereby, partial identity at least could be saved: "How far this soul finds a support for its existence beyond the body is: - another question. The everlasting nature of God, which in a sense is non-temporal and in another sense is temporal, may establish with the soul a peculiar intense relationship of mutual immanence. Thus in some important sense the existence of the soul may be freed from its complete dependence upon the bodily organization" (AI 208 italics added). Since the freedom from the body includes a freedom from the permanent threat of decay, it also holds the hope for diachronic identity for human persons:

The consequent nature of God is his judgment on the world. He saves the world as it passes into the immediacy of his own life. It is the judgment of a tenderness which *loses nothing that can be saved*. It is also the judgment of a wisdom which uses what in the temporal world is mere wreckage. Another image which is also required to understand his consequent nature is that of his infinite patience. ... God's role ... lies in the patient operation of the overpowering rationality of his conceptual harmonization. He does not create the world, he saves it: or, more accurately, he is the poet of the world, with tender patience leading it by his vision of truth, beauty, and goodness (PR 346 italics added).

So, in Whitehead's process-organismic approach, God may be able to save the Hobos from the "terror at the loss of the past, with its familiarities and its loved ones" (PR 340). They at least '[lose] nothing that can be saved'.

References

- Bennett J B (1973) Whitehead and Personal Identity. *The Thomist* 37(3): 510–521.
- Bradley F H (1922) *The Principles of Logic*. London: Oxford University Press.
- Butler J (1975) Of Personal Identity. In *Personal Identity*, ed. John Perry, 99–106. Berkeley: University of California Press [1736].
- Brand B L, Sar V, Stavropoulos P, Krüger C, Korzekwa M, Martínez-Taboas A (2016) Separating Fact from Fiction: An Empirical Examination of Six Myths About Dissociative Identity Disorder. *Harvard Review of Psychiatry* 24849. 257–270.
- Bröderbauer S, Huemer M, Riffert F (2013) On the Effectiveness of Incidental Hints in Problem Solving. *Gestalt Theory* 35(4): 349–364.
- Burkhardt H (1984) Identität/Differenz [Identity/Difference]. In *Lexikon der Erkenntnistheorie und Metaphysik*. [Encyclopedia of Epistemology and Metaphysics], ed. Friedo Ricken, 87–89. München: Beck.
- Carnap R (1961) *Der logische Aufbau der Welt*. [The Logical Structure of the World] Hamburg [1928].
- Caspar F (1995) *Plan Analysis – Toward Optimizing Psychotherapy*. New York: Winston.
- Caspar F (2007). Plan Analysis. In *Handbook of Psychotherapy Case Formulation* (2nd edition), ed. Tracy D. Eells, 251–289. New York: Guilford Press.
- Cobb J B (2007) *A Christian Natural Theology. Based on the Thought of Alfred North Whitehead*. Louisville: Westminster John Knox Press.
- Dewey J (1948) *Human Nature and Conduct*. New York.
- Eastman T (Ed.) (1997): Process Thought and Natural Science, I. Special Focus. *Process Studies* 26(3–4): 239–323.
- Eastman T (2004) Resource Guide for Physics and Whitehead. [Online Document] https://moam.info/resource-guide-for-physics-and-whitehead-cfcu_5a1943e81723dddc24762e7a.html
- Eastman T, Keeton H (Eds.) (2004) *Physics and Whitehead: Quantum, Process and Experience*. Albany: State University of New York Press.
- Eastman T, Riffert F (2009) Process Thought, Science and Philosophy. *World Futures. The Journal of General Evolution* (Special Issue on Process Philosophy) 65(1): 1–75.
- Edwards R B (1975) The Human Self: An Actual Entity or a Society? *Process Studies* 5(3): 195–203.

- Fetz R L (2019) *Die Wirklichkeit der Wirkwesen. Grundlegung einer organismischen und strukturalistischen Wirklichkeitskonzeption* [The Reality of Efficient Entities. On the Foundations of an Organismic and Genetic Structuralist Conception of Reality]. Freiburg: Alber.
- Fetz R L (1991) In Critique of Whitehead. *Process Studies* 20(1): 1–19.
- Fowler J W (1981) *Stages of Faith. The Psychology of Human Development and the Quest for Meaning*. San Francisco: Harper & Row.
- Fuchs T (2020) Embodiment and Personal Identity in Dementia. *Medicine, Health and Philosophy*. 23: 665–676.
- Fuhrmann M (1979) Persona, ein römischer Rollenbegriff [Persona, a Roman Role Concept] In *Identität* [Identity], eds. Odo Marquard and Karlheinz Stierle, 83–106.
- Galin D (2003) Untangling 'Self,' 'Person,' and 'I'. In *Searching for New Contrasts. Whiteheadian Contributions to Contemporary Challenges in Neuropsychology, Psychology, Psychotherapy, and the Philosophy of Mind*, eds. Franz Riffert and Michel Weber, 313–331. Vienna: Lang.
- Griffin D R (1998) *Unsnarling the World-Knot – Consciousness, Freedom and the Body-Mind Problem*. Berkley: University of California Press.
- Hartshorne Ch (1972a) The Compound Individual. In *Whitehead's Philosophy. Selected Essays, 1935–1970*, ed. Charles Hartshorne, 41–61. Lincoln: University of Nebraska Press.
- Hartshorne Ch (1972b) Personal Identity from A to Z. *Process Studies* 2(3), 209–215.
- Hume D (2007) *A Treatise of Human Nature*. Oxford: Oxford University Press [1739].
- Johnson A H (1951) *Whitehead's Theory of Reality*. Boston: Beacon Press.
- Kelly G A (1955) *The Psychology of Personal Constructs*. New York: Norton.
- Kohlberg L (1981) *Essays on Moral Development. Vol. I: The Philosophy of Moral Development*. San Francisco: Harper & Row.
- Kohlberg L (1984) *Essays on Moral Development. Vol. II: The Psychology of Moral Development*. San Francisco: Harper & Row.
- Lakoff G, Johnson M (1999) *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books.
- Langsam H (2001) Pain, Personal Identity, and the Deep Further Fact. *Erkenntnis* 54: 247–271.
- MacKay D (1987) Divided Brains – Divided Minds. In *Mindwaves: Thoughts on Intelligence, Identity & Consciousness*, eds. Colin Blackmore and Susan Greenfield, 5–16. London: Basil Blackwell.
- Madell G (1981) *The Identity of the Self*. Edinburgh: Edinburgh University Press.
- Mays W (1977) *Whitehead's Philosophy of Science and Metaphysics*. The Hague: Martinus Nijhoff.
- McCutcheon R A, Krystal J H, Howes O D (2020) Dopamine and Glutamate in Schizophrenia: Biology, Symptoms and Treatment. *World Psychiatry* 19(1): 15–33.
- Mulligan N W (2008) *Implicit Memory*. New York: Psychology Press.
- Nietzsche F (1980) Über Wahrheit und Lüge im außermoralischen Sinne [On Truth and Lies in an Extra-Moral Sense] (Bd.1). In *Friedrich Nietzsche: Sämtliche Werke*, eds. Giorgio Colli and Mazzino Montinari, 873–890. Berlin: De Gruyter [1873].
- Oser F, Gmünder P (1988) *Der Mensch – Stufen des religiösen Urteils. Ein strukturalistischer Ansatz* [Man – Levels of Her/His Religious Judgement.]. Gütersloh: Mohn.
- Parfit D (1986) *Reasons and Persons*. Oxford: Oxford University Press.
- Parfit D (1987) Divided Minds and the Nature of Persons. In *Mindwaves*, eds. Colin Blackmore and Susan Greenfield, 19–26. Oxford: Basil Blackwell.
- Perrig W J, Wippich W (1993) *Unbewusste Informationsverarbeitung* [Unconscious Information Processing]. Bern: Huber.
- Piaget J (1968) *Le structuralisme*. [The Structuralism] Paris: Presses Universitaires de France.
- Piaget J (1976a) *The Grasp of Consciousness*. Cambridge: Harvard University Press.
- Piaget J (1976b) *Die Äquilibration der kognitiven Strukturen* [The Equilibration of Cognitive Structures]. Stuttgart: Klett.
- Piaget J (1980) *Experiments in Contradiction*. Chicago: Chicago University Press.
- Putnam F W (1991) Recent Research on Multiple Personality Disorder. *Psychiatric Clinics of North America* 14(3): 489–502.
- Reid T (1975) Of Mr. Locke's Account of Our Personal Identity. In *Personal Identity*, ed. John Perry, 33–52. Berkley: University of California Press. [1785]
- Rescher N (1996) *Process Metaphysics: An Introduction to Process Philosophy*. New York: State University of New York Press.
- Rescher N, Urquhart A (1971) *Temporal Logic*. New York: Springer.
- Riffert F (1995) *Whitehead und Piaget. Zur interdisziplinären Relevanz der Prozessphilosophie*

[Whitehead and Piaget. *On the Interdisciplinary Relevance of Process Philosophy*]. Vienna: Lang.

Riffert F (2003a) On Scientific Confirmation of Causal Efficacy. In *Searching for New Contrasts. Whiteheadian Contributions to Contemporary Challenges in Neurophysiology, Psychology, Psychotherapy, and the Philosophy of Mind*, eds. Franz Riffert and Michel Weber, 171–180. New York: Lang.

Riffert F (2003b) Whiteheads Ethik [„Whitehead’s Ethics“]. *Freiburger Zeitschrift für Philosophie und Theologie* 50(1,2): 171–187.

Riffert F (2003c) On Non-Substantialism in Psychology. *International Journal for Field-Being* (online journal). Special Issue: Whitehead and the Sciences.

Riffert F (2004a) Whitehead’s Process Philosophy as Scientific Metaphysics. In *Physics and Whitehead: Quantum, Process and Experience*, eds. Timothy Eastman and Hank Keeton, 199–222. Albany: State University of New York Press.

Riffert F (2004b) Whitehead sur la métaphysique, la science et les mathématiques. [Whitehead on Metaphysics, Science and Mathematics]. In *Alfred North Whitehead. De l’algèbre universelle à la théologie naturelle. [Alfred North Whitehead. from Universal Algebra to natural Theology]*, eds. Françoise Beets, Michel Dupuis and Michel Weber, 149–169. Frankfurt: ontos.

Riffert F (2007b) Whitehead and Piaget on the Construction of Permanent Objects. In *Perception Reconsidered. The Process Point of View*, ed. Franz Riffert, 175–187. New York: Lang.

Riffert F (2019) Consciousness: The Point of View of Process Philosophy and Genetic Structuralism: A Critical Comparison and Some Consequences. *Balkan Journal of Philosophy* 11(2): 83–106.

Riffert F, Cobb J B (2003) Reconnecting Science and Metaphysics: General Considerations and Pioneer Works on Process Psychology. In *Searching for New Contrasts. Whiteheadian Contributions to Contemporary Challenges in Neurophysiology, Psychology, Psychotherapy, and the Philosophy of Mind*, eds. Franz Riffert and Michel Weber, 19–35. New York: Lang.

Riffert F, Eastman T (2008) The Role of Process Philosophy in a World of Science. *World Futures: The Journal of General Evolution* 64(2) (Special Issue): 73–149.

Riffert F, Sander H J (Eds.) (2008) *Researching with Whitehead: System and Adventure*. Freiburg i. Br.: Alber.

Riffert F, Weber M (2003) *Searching for New Contrasts. Whiteheadian Contributions to Contemporary Challenges in Neurophysiology, Psychology, Psychotherapy, and the Philosophy of Mind*. New York: Lang.

Riffert F, Huemer M, Bröderbauer S (2019) Whitehead, Microgenesis, and the Unconscious. In *Science and Mind in Contemporary Process Thought*, eds. Jakub Dziadkowiec and Lukasz Lamza, 72–88. Newcastle: Cambridge Scholar Publishing.

Salmon W C (1983) *Logik [Logic]*. Stuttgart: Reclam.

Schacter D L (1987) Implicit Memory: History and Current Status. *Journal of Experimental Psychology* 13(3): 501–518.

Scheer J W, Catina A (1993) *Einführung in die Repertory Grid-Technik [Introduction into the Grid-Technique]*. Bern: Huber.

Süllwold L (1983) *Schizophrenie [Schizophrenia]*. Berlin: Kohlhammer.

Weiskrantz L (1997) *Consciousness Lost and Found: A Neuropsychological Elaboration*. Oxford: Oxford University Press.

Whitehead A N (1898) *A Treatise on Universal Algebra with Applications*. Cambridge: Cambridge University Press. (UA)

Whitehead A N (1906) *The Axioms of Projective Geometry*. London: Cambridge University Press. (APG)

Whitehead A N (1911) *An Introduction to Mathematics*. Williams & Norgate. (IM)

Whitehead A N (1967) *Science and the Modern World*. New York: Macmillan [1925]. (SMW)

Whitehead A N (1926) *Religion in the Making*. New York: Macmillan. (RM)

Whitehead A N (1978) *Process and Reality*. New York: Free Press [1929a]. (PR)

Whitehead A N (1958) *The Function of Reason*. Princeton: Princeton University Press [1929b]. (FR)

Whitehead A N (1967) *Adventures of Ideas*. New York: Macmillan [1933]. (AI)

Whitehead A N (1968) *Modes of Thought*. New York: Macmillan [1938]. (MT)

Whitehead A N (1947) *Essays in Science and Philosophy*. New York: Philosophical Library. (ESP)

Invited paper

TRUTH AND MODALITIES (I)*

Fabrice Pataut

CNRS, UMR 8011 Sciences, Normes, Démocratie
Paris, France
email: patautfabrice@gmail.com

Abstract:

I contrast two construals of the thesis that truth is independent of verifiability in principle: a modal one and a non modal one. I argue in favor of the modal construal and then, on that basis, that independence holds across the board, i.e., even for statements that are verifiable by us relative to familiar, customary, non-skeptical standards.

Key words: Bivalence, Contingency, Disquotation, Knowability, Modalities, Necessity, Possibility, Semantic Realism, Truth ; Bernays (Paul), Dummett (Michael A. E., Sir), Hilbert (David), Loar (Brian), Peirce (Charles Sanders).

1. Realism and independence
2. Possibilities
3. Contingencies
4. Conclusions

References

1. Realism and independence

In what might turn out to be the fundamental sense of the word “realist,” inherited from Michael Dummett, a realist about a given class of statements holds that the statements of that class could

be true even though we ignore that they are. Their truth, the realist claims, is independent of our ability to recognize the fact by the appropriate means (see, e.g., *Dummett [1973] 1978*).

This formulation of the doctrine raises a crucial issue for both realists and their opponents. Should the realist be ready to argue for the independence claim and the antirealist for its rejection even when we happen to be able to acknowledge the truth of the statements of some disputed class? Or, on the contrary, should the realist restrict its scope so that the claim applies to *pro tempore* undecided statements only, and the antirealist rest content with it as far as statements that are verifiable by normal, customary, non-skeptical standards are concerned? Dummett and most of his opponents share the conception that the realism vs. antirealism debate exists for *pro tempore* undecided statements only and that decidable or verifiable statements are immune to the disagreement, so that whenever, e.g., proofs are available in mathematics, or data (behavioral or otherwise) may be gathered for statements about the mental states of others, or evidence pro or con may be recovered now or at some later time for statements about the past, the statements belonging to these particular classes are immune to the disagreement. This, I shall argue, is mistaken.

If the modal construal of the realist doctrine is chosen, the realist is de facto committed to the view

* The article will be published in three separate installments. The first includes section 1 only.

that the statements of some disputed class, if true, are so whether or not we are in a position to acknowledge their truth or are able to put ourselves in that position. At first blush, it seems that this is what independence means, implies or is about, and that the realist's contentious claim must accordingly be construed as holding across the board, i.e., for classes of decidable statements and for classes of *pro tempore* undecided ones alike without discrimination. The claim to the effect that the two cases are substantially dissimilar would be invidious and the reason for this lies in our understanding of the notion of independence. Independence is a matter of autonomy or self-sufficiency. For a thing *A* (say, truth) and another thing *B* (say, its recognizability by us) to be independent from each other, it must be the case that *A* could obtain, or be instantiated, or make sense, without *B* also obtaining, or being instantiated, or making sense.

In the case we're considering here, it isn't so obvious the converse should hold. It might seem on the contrary that if *B* obtains, *A* thereby does simply because when the appropriate recognitional abilities with respect to truth are correctly activated, or could be, truth is thereby secured, or could be (provided skepticism isn't an option). If it turns out that some statement is true once our inquiry has reached completion, or would be should we decide to inquire, then its truth *isn't* independent from its actual recognition or in principle recognizability by us (at least not anymore). In other words, one might complain that whenever the appropriate capacities are successfully activated, or may be, *dependence*, rather than independence, is our lot.

Now if one is generous and lenient rather than strict and inflexible with respect to the notion that should replace truth or constrain it in the antirealist perspective, a different conclusion might be drawn. If one takes truth to be truth à la Peirce in the long run of scientific discovery (see *Peirce [1905] 1998*), it still might be argued that since *B* doesn't yet obtain and will obtain only in an idealized epistemic state we still are very far from, *A* is, in a perfectly legitimate sense, independent from *B*. Only a strict construal of the antirealist position according to which truth is to be acknowledged in a finite number of steps, say by way of verification in polynomial time so that the feasibility of the verification is guaranteed, would warrant the conclusion to the effect that the converse also holds.

Although the disagreement about the weak vs.

the exacting construals is a genuine one as far as the strength of the position to be defended by Dummett's antirealist is concerned, it would be a mistake to argue that the dissent threatens in any way the workableness or intelligibility of the opponent's position. A realist defending the modal claim isn't disputing the truth of actually verified statements (whether or not the verification is carried out in polynomial time, or in some non polynomial yet suitably short time). The crucial point, rather, is that statements that are either conclusively verifiable in the long run of our inquiry, or actually verified by us in some specified short span of time, *logically* could be true even though not (as in fact they are) either verifiable in principle or actually verified by us *hic et nunc* (see, e.g., *Loar 1987: 85*)¹.

If that is so, then it would seem that for any *s* of some disputed class:

$$(1) \Diamond (s \text{ is true} \bullet s \text{ is unverifiable by us})$$

holds just because there is no link or connection, either conceptual or practical between truth and its recognition. The two are mutually independent in two different albeit related ways: in the sense that no notion or *concept* of recognition, either feasible or in principle, may be extracted from the notion of truth, as if were analytically, and in the sense that no *practical* use or exercise of recognitional capacities fitted for *s*'s area of discourse may constrain truth (or allow us to eschew it altogether).

Several points might be at stake here, related either to concepts or practices, and we must accordingly distinguish between them.

First, there is the claim that no inconsistency, or logical contradiction, or paradox, is implied in our holding (1) or follows from it: (1) is consistent with the laws of logic and nothing downright illogical follows, either from its assertion or from what logically follows from it.²

¹ Whether or not this also commits us to the view that truth is a substantial property, over and above disquotation, is a further problem I shall consider later on.

² There is, of course, a disagreement regarding which laws have to be counted as the laws of logic, so that one might object that what is consistent with the laws of *some* logic, or what logically follows from them, turns out to be inconsistent with the laws of *another one* and does not follow at all from its laws. I shall go back to this crucial point of this particular formulation or aspect of the independence claim later on.

Secondly, there is the contention that no entrenched cognitive shortcoming makes it impossible for us to grasp one (the truth of s) without thereby grasping the other (the verifiability of s by us).³ This may be construed as a negative claim about concept formation: there is no reason to believe that we cannot form a bona fide concept of truth *not* constrained by whatever appropriate epistemic notion would constrain it in the antirealist perspective, given the area of discourse the statement belongs to, e.g., proof in the mathematical case.

One may also give a Kantian flavour to the debate and turn it into a debate over whether or not it is necessary for us to conceive of truth, or to represent truth as being constitutively linked to the conditions of its recognizability by us. Under that construal of the realist proposal, the advocate of (1) would be opposing a Kantian or neo-Kantian point to the effect that in order for truth to be a proper object of cognition, some a priori conditions have to be met regarding its accessibility to us. In yet other words, the advocate of (1) would be rejecting the view that in order for the truth of any s to be grasped at all, the following epistemic necessity should hold:

(2) $\Box (s \text{ is true} \rightarrow s \text{ is verifiable by us})$.

The modal notion to be rejected by the realist would have to be epistemic because the necessity at play here would have to be unpacked as a notion that embeds the conditions that have to be satisfied or fulfilled in order for truth to be known or grasped at all *by us* (as opposed to a notion that would have to be grasped by other beings with a different cognitive set up), so that a bona fide notion of truth must abide by (2).

One might want to recall here the Kantian distinction between the conceptual properties that objects of kind X possess in general (as opposed to the individual properties that some particular in-

stances of X don't share with other particulars of the same kind) and the *intuitional properties* that all X possess insofar as they are objects of our intuition. The Kantian point is that these so-called intuitional properties may not be extracted, either from the general concept or from the individual instances. The neo-Kantian point would be that for the truth of any s to be grasped or acknowledged at all, the truth-conditions of s must possess properties that allow these conditions to be the object of our cognition (whether or not of an intuitional kind). So there is no choice: the modal operator $\Box$ in (2) *must be read as epistemic*. In yet other words, given the relevant information we are able to gather about how the world is, i.e., given the way the truth-conditions of our statements about the world are fulfilled when they are, our claims regarding these conditions must be construed as being about how things must be given our information about *how the world in fact is*, rather than claims about how things must be in *all the various ways the world could have been*.

There are of course crucial differences between these standpoints, to which I shall return. For the time being one might perhaps unpack the realist standpoint at stake in the debate in the following way. One difficulty in this respect is due to the difference between the view according to which the concept of verifiability in principle, or of actual verification, is already included in the notion of truth, and the view according to which the capacities that have to be activated in order to establish truth constrain the semantic notion. The first one might be called a *containment* view, the second, a *recognitional* view.

Independence with respect to the possibility of truth determination doesn't as it were vanish, when, as a matter of contingency, the possibility becomes actualized, i.e., when we happen to possess the abilities which must be activated in order for us to decide the truth of statements of a given class and when we succeed in carrying out proofs, or in gathering empirical evidence. The possibility of recognition, verification or acknowledgement is a genuine or natural one (provided that we disregard skeptical standards) and there is no reason to think that it may (or indeed must) always be activated, or that we can't make sense of evidential relations between us and the truth of some s breaking down so that s 's truth conditions are divorced from its assertability conditions.

³. This formulation is the one closest to Dummett's concern. Dummett's antirealist claims that our understanding of a statement, or of a statement's meaning consists in a mastery of its verifiability conditions, so that we have no bona fide grasp or understanding of its verification-transcendent truth conditions, i.e., of what it would be like for that statement to be true independently of our possibly acknowledging its truth. See, e.g., Dummett [1973] 1978 and Loar 1987, especially pages 85–86.

If that line of argument is correct, the realist is indeed committed to the view that what makes the statements of the disputed class true is something which is not intrinsically connected to the means of recognition we must appeal to when acknowledging truth, i.e., proofs in the mathematical case and the identification and gathering of defeasible, partial and gradual evidence in the empirical case.

This lack of constitutive or intrinsic connection between the truth-makers of statements and our human capacities of recognition might perhaps be illustrated by the platonist standpoint in the philosophy of mathematics when, e.g., mathematical statements of an axiomatized theory are taken as privileged examples.

Paul Bernays, comparing Euclid's axiom system to Hilbert's, writes:

[...] Euclid speaks of figures to be *constructed*,* whereas, for Hilbert, system of points, straight lines, and planes exist from the outset. Euclid postulates: One can join two points by a straight line; Hilbert states the axiom: Given any two points, there exists a straight line on which both are situated. "Exists" refers here to existence in the system of straight lines.

This example shows already that the tendency of which we are speaking consists in viewing the objects as cut off from all links with the reflecting subject.

Since this tendency asserted itself especially in the philosophy of Plato, allow me to call it "platonism."

* The italics are provided by Charles Parsons in his translation from the French.

Paul Bernays [1935] 1983: 258–259

Should anyone holding (1) and rejecting (2) follow that tendency and view both the objects and the statements about them as cut off from all links with the abilities we have at our disposal in our capacity as reflecting subjects?

Dummett's antirealist quite naturally resists the inclination Bernays writes about here. The question we're asking now is whether his opponent should be in agreement with him when considering classes of decidable or verifiable statements.

Dummett's views on meaning strongly suggest that he should. This is where what I've called the recognitional view, as opposed to the containment view, comes into the picture. If, as Dummett's con-

tends, our understanding of the meaning of any statement whatsoever, either decidable or not, consists in a mastery of its verifiability conditions, it follows that we cannot have any mastery or bona fide understanding of what it would be like for a statement to be both true and unverifiable by us. There is just no way for us to master such conditions, or indeed form in any way a bona fide concept of them.

Part of the explanation of why a Dummettian antirealist will reject the modal claim under *any* construal comes from the way the satisfiability of the manifestability requirement for the decidable cases is construed. We may agree with Dummett's antirealist that when the statements are decidable, our grasp or mastery of their verifiability conditions is exhaustively manifested by our actually deciding the statements or by showing that we are in a position to do so. This is obviously not something we may do when the statements are *pro tempore* undecidable. The knowledge that a statements's truth-conditions are satisfied, no matter how it may be manifested, cannot be equivalent to the knowledge that the very same conditions could be fulfilled unbeknownst to us.

In yet other words, we cannot achieve a mastery of the verification-conditions of either (1), or of the non-modal

(3) (s is true • s is unverifiable by us).

The proposal, as offered by Dummett in *Dummett [1973] 1978*, is nevertheless unsatisfactory as far as independence is concerned for the following reason.

The observation that the first knowledge or mastery may be manifested in a certain way gives us no clue as to how the second might be manifested. If we unpack the contrast put before us at the disadvantage of the *pro tempore* undecided cases, what we are being offered is a remark to the effect that a knowledge or a mastery of the truth-conditions of particular instances of "it is possible that s is true although we ignore it" is not something that may be construed on a par with a knowledge or a mastery of the truth-conditions of instances of s when s is decidable. It cannot indeed because if we were to look for a justification for our inability to verify s , we would have to begin by attempting to verify s . Our failure to decide the *pro tempore* undecided s should lead us to conclude

that the negation of *s* holds. Once the negation of *s* is established, the road to an establishment of the independence of the truth of *s* is closed. It isn't possible to show that *s* is absolutely neither provable nor refutable: intuitionistic negation forbids this.⁴ In other words, there is no way of manifesting a mastery or knowledge of the truth-conditions of instances of the modal claim that could satisfy Dummett's antirealist constraints on truth given that the intuitionistic reading of negation is presupposed (as opposed to justified).

Another important point must be stressed concerning the neo-Kantian construal of (2) in relation to the logical constant for negation. What is claimed here is that the only bona fide notion of "the truth of *s*" we may grasp, for any *s*, is one according to which these conditions are constrained by knowledge. There are no "things in themselves" embedded in such conditions. We might perhaps be able to think of them somehow as things in themselves, but we cannot *cognize them* as things in themselves (see Kant [1781, 1787] 1997: A35/B51, B69, B36, A360 and Allison 2004). Just as we credit ourselves with a grasp of the meaning of a classical constant that we, as a matter of fact, can't achieve when making inferences using classical negation, e.g., when eliminating double negation, we credit ourselves with a knowledge we do not possess when we credit ourselves with a grasp of the possibly verification-transcendent truth-conditions of our statements, whether atomic or complex (i.e., built up from atomic statements by a finite application of rules governing the logical constants). We might be able to think or to represent such meaning and conditions as cut off from the way we fix the meaning and grasp the conditions, but the thought or representation thereby obtained is *without* any genuine cognitive content.

References

Allison H E (2004) *Kant's Transcendental Idealism – An Interpretation and Defence*, Revised and Enlarged Edition. New Haven: Yale University Press.

Descartes R [1641] (1967) *Meditations on First*

Philosophy. In *Philosophical Works of Descartes*, vol. I, eds E. S. Haldane and G. R. T. Ross, 1–32, Cambridge, Cambridge University Press.

Dummett M A E [1973] (1978) The Philosophical Basis of Intuitionistic Logic. In *Truth and other Enigmas*, 1st ed., 215–247. Cambridge, Mass: Harvard University Press.

Edgington D (1985) The Paradox of Knowability. *Mind* 94(376): 557–568.

Horwich Paul (1998) *Truth*, 1st ed., Oxford: Basil Blackwell.

James W (1907) Pragmatism's Conception of Truth, *The Journal of Philosophy, Psychology and Scientific Methods* 4(6): 141–155.

Kant I [1781, 1787] (1997) *Critique of Pure Reason*, translated and edited by P. Guyer and A. Wood, Cambridge: Cambridge University Press.

Kripke S [1975] (1984) Outline of a theory of truth. In *Recent Essay on Truth and the Liar Paradox*, ed. R. L. Martin, 53–81. Oxford, Oxford University Press.

Loar B (1987) Truth Beyond All Verification. In *Michael Dummett: Contributions to Philosophy*, ed. B. Taylor, 81–116. Dordrecht: Martinus Nijhoff Publishers.

Martin R L ed. (1984) *Recent Essays on Truth and the Liar Paradox*. Oxford: Clarendon Press.

Martin R L, Woodruff P W [1975] (1984) On Representing 'True-in-L' in *L*. In *Recent Essays on Truth and the Liar paradox*, ed. R. L. Martin, 47–51. Oxford: Oxford University Press

McDowell J (1976) Truth Conditions, Bivalence and Verificationism. In *Truth and Meaning – Essays in Semantics*, eds G. Evans and J. McDowell, 42–66. Oxford: Clarendon Press.

Montgomery H A, Routley F R (1966) Contingency and non-contingency bases for normal modal logics. *Logique et Analyse* 9(35–36): 318–328.

Parsons C [1974] (1984) The Liar Paradox. In *Recent Essays on Truth and the Liar paradox*, ed. R. L. Martin, 9–45. Oxford: Oxford University Press.

Peirce C S [1905] (1998) What Pragmatism Is. In *The Essential Peirce*, vol. 2, The Peirce Edition Project, 331–345. Bloomington, Indiana: Indiana University Press.

Putnam H (1981) *Reason, Truth and History*, Cambridge: Cambridge University Press.

Williamson T (1987), On the Paradox of Knowability. *Mind* new series 96(382): 256–261.

⁴ See Dummett 1977: 17.

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

Wright C (1998) Truth : A Traditional Debate Reviewed. *Canadian Journal of Philosophy*, 24: 31–74.

A BILATERAL REDUCTION SENTENCE FOR MODULATION*

Miguel López-Astorga

Institute of Humanistic Studies “Juan Ignacio Molina”,
University of Talca, Talca Campus, Chile
email: milopez@utalca.cl
ORCID ID: <https://orcid.org/0000-0002-6004-0587>

Abstract:

The theory of mental models considers sentential connectives to refer to semantic possibilities. The theory also proposes that the possibilities can be amended by virtue of modulation processes, that is, processes in which semantics or pragmatics can have an influence. The definition of modulation the theory of mental models gives has been deemed as a scientific definition within Carnap's framework. This is because it is easy to find reduction sentences corresponding to it. What seems to be harder is to think about bilateral reduction sentences for that definition. This paper addresses this issue. It shows that at least a bilateral reduction sentence is possible for modulation.

Key words: bilateral reduction sentence; modulation; possibility; reduction sentence; theory of mental models.

Introduction

A very important concept in cognitive science in general and in the theory of mental models in particular is that of modulation (e.g., Quelhas &

Johnson-Laird, 2017; Quelhas, Johnson-Laird, & Juhos, 2010). According to that theory, the interpretation of language consists of the identification of semantic possibilities or models associated with each sentence (see also, e.g., Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). But semantics and pragmatics can alter the initial possibilities of a sentence. This can happen because of modulation. Modulation can adjust and even remove possibilities (see also, e.g., Byrne & Johnson-Laird, 2020).

The general definition of modulation the theory of mental models offers is as follows:

“The process in the construction of models in which content, context, or knowledge can prevent the construction of a model and can add information to a model” (Johnson-Laird, Khemlani, & Goodwin, 2015, p. 202).

Checking whether or not (1) is a scientific definition has been attempted. To do that, Carnap's (1936, 1937) general framework has been assumed. The result has been positive: reduction sentences (in particular, three reduction sentences) can be built for (1) (López-Astorga, 2022).

The issue that the present paper addresses is the fact that Carnap (1936) not only speaks about reduction sentences, but also deals with reduction pairs and bilateral reduction sentences. The three reduction sentences proposed in the literature for modulation are not problematic. But it is important

* PIA Ciencias Cognitivas, Centro de Investigación en Ciencias Cognitivas, Instituto de Estudios Humanísticos, Universidad de Talca. Fondo Fondecyt de Continuidad para Investigadores Senior, código FCSN2102, Universidad de Talca.

to analyze whether or not the essential concepts involved in (1), that is, content, context, and knowledge, allow constructing reduction pairs and bilateral reduction sentences too. The aim here is to show that this is possible. Therefore, the paper tries to argue that the definition of modulation in the theory of mental models is scientific in accordance with Carnap's (1936, 1937) different criteria.

To achieve that goal, this paper will have several sections. One of them will describe the way Carnap (1936) considers scientific definitions. This will be explained by means of the concepts of reduction sentence, reduction pair, and bilateral reduction sentence. Then, how the theory of mental models understands modulation will be accounted for. The next section will show the reduction sentences that have been proposed for modulation. One more section will point out the objections that can be offered for those sentences based on the manner in which Carnap (1936) presents reduction pairs and bilateral reduction sentences. Finally, how the objections can be overcome will be indicated. This will allow for the claim that the definition of modulation in the theory of mental models is scientific and progressively confirmable in an even stronger sense than that sustained in the literature.

Rudolf Carnap and scientific definitions

Carnap (1936) seems to agree with Popper (2002) about the impossibility of absolutely verifying sentences. For that reason, Carnap (1936) presents reduction sentences, reduction pairs, and bilateral reduction sentences in order to gradually confirm whether or not particular properties can be associated with particular definitions. Thus, to know the properties that can correspond to a predicate R, a sentence such as (2) can be built.

$$\forall x [Px \rightarrow (Qx \rightarrow Rx)],$$

where '∀' is the universal quantifier, P and Q are predicates, and '→' represents a conditional relation.

According to Carnap (1936), (2) is 'a reduction sentence for R', since it enables one to check to what extent predicates such as P and Q can be linked to R. But this can be only if requirement (3) is fulfilled.

$$\neg \forall x \neg (Px \wedge Qx),$$

where '¬' is negation and '∧' stands for conjunction.

Requirement (3) provides that cases of P and Q at the same time have to exist.

Reduction sentence (2) is not the only kind of sentence that can be constructed to know what the

properties that can be attributed to R are. Following Carnap (1936), it is also possible to build a reduction pair such as that consisting of (2) and (4).

$$\forall x [Sx \rightarrow (Tx \rightarrow \neg Rx)],$$

where S and T are predicates.

Sentences (2) and (4) are 'a reduction pair for R'. This reduction pair not only allows one to review whether or not P and Q can be related to R. It also enables one to analyze whether or not S and T are properties incompatible with R as well. Carnap (1936) provides a requirement for reduction pairs too:

$$\neg \forall x \neg [(Px \wedge Qx) \vee (Sx \wedge Tx)],$$

where '∨' indicates disjunction.

Requirement (5) claims that either at least a case of P and Q at once or a case of S and T at once should exist.

Lastly, a bilateral reduction sentence such as (6) can also be built.

$$\forall x [Px \rightarrow (Qx \leftrightarrow Rx)],$$

where '↔' expresses a biconditional relation.

Sentence (6) is 'a bilateral reduction sentence for R'. What (6) tries to confirm is whether or not the link between Q and R is much stronger than in (2). Carnap (1936) gives a requirement here as well:

$$\neg \forall x \neg Px.$$

Requirement (7) points out that at a minimum a case of P is necessary.

Reduction sentences have been offered for definition (1). But it appears that the construction of reduction pairs or bilateral reduction sentences from those reduction sentences can be controversial. This issue will be developed below. An important point in the argumentation will be related to the fact that, as Carnap (1936) acknowledges, (6) gets derived from (2) and (4) if (8) is the case.

$$\forall x [(Px \leftrightarrow Sx) \wedge (\neg Qx \leftrightarrow Tx)].$$

Sentence (8) serves to transform (4) into (9).

$$\forall x [Px \rightarrow (\neg Qx \rightarrow \neg Rx)]$$

and (6) can be inferred from (2) and (9).

This point will be essential for the objection that can be proposed regarding the three reduction sentences that have been offered for (1). Before showing this, the next section explains how modulation works in the theory of mental models.

Modulation in the theory of mental models

According to the theory of mental models, modulation can have an influence on any sentential connective (see also, e.g., Johnson-Laird, 2012). But given that the example used in the literature to present the reduction sentences for (1) is disjunction (López-Astorga, 2022), the present paper will only consider, as an instance, this last connective. However, it is important not to forget that the account would be similar for any other sentential connective.

In principle, a disjunction such as (10) would have three possibilities or models (see also, e.g., Quelhas, Rasga, & Johnson-Laird, 2019).

Either they eat salad or they eat fries

The possibilities or models would represent situations in which (10) would be true (see also, e.g., Johnson-Laird & Ragni, 2019). Those possibilities or models would be (11), (12), and (13).

It is possible that they eat salad and they eat fries

It is possible that they eat salad and they do not eat fries

It is possible that they do not eat salad and they eat fries

Although cases (11), (12), and (13) seem to reveal that the theory of mental models is not different from logic, it is. Many features distinguish the theory of mental models from logic (see also, e.g., Espino, Byrne, & Johnson-Laird, 2020; Johnson-Laird, 2010; Khemlani, Hinterecker, & Johnson-Laird, 2017; López-Astorga, Ragni, & Johnson-Laird, 2022). To explain all of them would move this paper away from its aim. Only one characteristic differentiating the theory from logic is important here: modulation.

Modulation in the theory of mental models is defined as (1). Hence, it can happen by virtue of content, context, and knowledge. In other words, content, context, and knowledge can modify models or possibilities. (The rest of this section is based on the account, the examples, and the models or possibilities assigned to those examples in López-Astorga, 2022.)

The case of content can be the following:

“Either you eat rice or you eat rice and chicken”
(López-Astorga, 2022, p. 112).

While it is a disjunction, sentence (14) only enables two possibilities: (15) and (16).

It is possible that you eat rice and you do not eat chicken

It is possible that you eat rice and you eat chicken

If (14) is true, there is no other possibility. This fact has to do with content. In any context in which sentence (14) is said, people can note that (15) and (16) are the only possibilities, even if they do not know what rice and chicken are. The word ‘rice’ appears in the two disjuncts. That implies that to know its meaning is not necessary in order to be aware that the situation in which you eat rice is always the case. So, neither context nor knowledge plays a role here.

As far as context is concerned, an example can be that of the following circumstance (taken from López-Astorga, 2022): two people read a menu indicating that they can choose an apple or an orange for dessert. If after eating an apple, they ask whether they can also take an orange, what (17) means is that they cannot.

“Either you eat an apple or you eat an orange”
(López-Astorga, 2022, p. 114).

In the context described, given that they have already eaten an apple, the only model or possibility that can be attributed to (17) is (18).

It is possible that you eat an apple and you do not eat an orange

This context restricts the models or possibilities. The content would enable more models or possibilities without it. Regarding knowledge, its action is not present in this case either. What (17) implies as an answer in the context is not directly related to the meaning of the words ‘apple’ and ‘orange’. If the two alternatives for dessert were different, the situation would not be distinct.

Finally, the example for knowledge can be (19).

“Either Pat visited Milan or she visited Italy”
(Johnson-Laird et al., 2015, p. 204; see also López-Astorga, 2022, p. 108).

Now, the models or possibilities are (20) and (21).

It is possible that Pat visited Milan and she visited Italy

It is possible that Pat did not visit Milan and she visited Italy

Pat cannot be in Milan without being in Italy.

So, there are no more possibilities. And this is only because of knowledge. The possibilities for (19) are (20) and (21) in any context (19) is said. On the other hand, content does not have an influence in the sense it has in (14).

From these examples, reduction sentences can be offered. The next section will show that. But it is important to remember that only content causes modulation in (14) (neither context nor knowledge), only context causes modulation in (17) (neither content nor knowledge), and only knowledge causes modulation in (19) (neither content nor context).

Modulation and its reduction sentences

Given (1) and the fact that content, context, and knowledge can, separately, modulate sentences, to construct reduction sentences corresponding to the definition of modulation is easy. (The rest of this section will also follow López-Astorga's, 2022, account and its reduction sentences.)

It is enough to assume these equivalences:

P: To be a sentence

R: To be modulated

U: To have content modifying models or possibilities

V: To be said in a context modifying models or possibilities

W: To be related to knowledge modifying models or possibilities

In this way, these three reduction sentences can be built:

$$\forall x [Px \rightarrow (Ux \rightarrow Rx)]$$

$$\forall x [Px \rightarrow (Vx \rightarrow Rx)]$$

$$\forall x [Px \rightarrow (Wx \rightarrow Rx)]$$

Formulae (22), (23), and (24) capture the three cases in which modulation can happen. For this reason, it can be said that they are reduction sentences for R. Even (25) could be claimed.

$$\forall x [Rx \leftrightarrow (Ux \vee Vx \vee Wx)]$$

This is essentially the account that can be found in the literature (López-Astorga, 2022). It seems to reveal that the definition of modulation is coherent with the requirements Carnap (1936, 1937) provides to construct scientific definitions that can be gradually confirmed. But, if reduction pairs and bilateral reduction sentences are considered, all of

this appears to be more complex. The reason will be shown below.

Are the reduction sentences for modulation incompatible?

The main difficulty (22), (23), and (24) present is that they can include mutually exclusive predicates. (22) is the reduction sentence corresponding to (14). (14) is a sentence in which the action of modulation is just by virtue of content (with no role for context or knowledge). That implies (26).

$$\exists x (Px \wedge Ux \wedge \neg Vx \wedge \neg Wx),$$

where '∃' is the existential quantifier.

Formulae (22) and (26) suffice to lead to Rx. But this does not allow one to construct a bilateral reduction sentence. (22) and (27) cannot be a reduction pair for R enabling the derivation of (28).

$$\forall x [Px \rightarrow (\neg Ux \rightarrow \neg Rx)]$$

$$\forall x [Px \rightarrow (Ux \leftrightarrow Rx)]$$

Sentences (23) and (24) show the falsity of (27) and, accordingly, (28).

Something similar occurs with (23). It is the reduction sentence corresponding to (17). In (17), context is the only component acting for modulation. Hence, given (17), (29) can be accepted.

$$\exists x (Px \wedge Vx \wedge \neg Ux \wedge \neg Wx)$$

Again, (23) and (29) are enough to conclude Rx. But a bilateral reduction sentence is hard here too. (23) and (30) are not a reduction pair for R.

$$\forall x [Px \rightarrow (\neg Vx \rightarrow \neg Rx)]$$

As in the previous case, (22) and (24) prevent (30) and (31) from being true.

$$\forall x [Px \rightarrow (Vx \leftrightarrow Rx)]$$

And the same happens with (24). This last sentence can be linked to (19). Knowledge is what modulates in (19), without influence from content or context. Therefore, (32) should be assumed.

$$\exists x (Px \wedge Wx \wedge \neg Ux \wedge \neg Vx)$$

Although (24) and (32) allow one to deduce Rx, once again, a bilateral reduction sentence is difficult. (24) and (33) cannot be accepted as a reduction pair for R.

$$\forall x [Px \rightarrow (\neg Wx \rightarrow \neg Rx)]$$

This is because (22) and (23) show that (33) and (34) are false.

$$\forall x [Px \rightarrow (Wx \leftrightarrow Rx)]$$

So, it can be stated that, given the definition of modulation in (1), it is possible to provide reduction sentences such as (22), (23), and (24). Nevertheless, this seems to be different in the case of bilateral reduction sentences. It appears that bilateral reduction sentences linked to (1) are hard. If this is the case, the definition of modulation the theory of mental models gives is weak within Carnap's (1936, 1937) framework, at least in a sense: it does not allow for reference to bilateral relations. But this is not correct. The next section tries to show that.

Building a bilateral reduction sentence for modulation

Although it may seem to be the opposite, reduction pairs are possible as well. For example, (4) could be taken along with (22), (23), or (24) if (35) holds.

$$\forall x \{ (Px \leftrightarrow Sx) \wedge [Tx \leftrightarrow (\neg Ux \wedge \neg Vx \wedge \neg Wx)] \}$$

To consider predicates U, V, and W at once reveals that a bilateral reduction sentence is also possible. (25) gives an important clue too. By virtue of (35), (4) can be transformed into (36).

$$\forall x \{ Px \rightarrow [(\neg Ux \wedge \neg Vx \wedge \neg Wx) \rightarrow \neg Rx] \}$$

and (36) is equivalent to (37).

$$\forall x \{ Px \rightarrow [\neg(Ux \vee Vx \vee Wx) \rightarrow \neg Rx] \}$$

Therefore,

$$\forall x \{ Px \rightarrow [Rx \rightarrow (Ux \vee Vx \vee Wx)] \}$$

On the other hand, if, for example, (22) is true, (39) is too.

$$\forall x \{ Px \rightarrow [(Ux \vee Vx \vee Wx) \rightarrow Rx] \}$$

In fact, (39) is also true in the cases in which (23) and (24) are, respectively, true. Hence, (38) and (39) lead to bilateral reduction sentence (40).

$$\forall x \{ Px \rightarrow [(Ux \vee Vx \vee Wx) \leftrightarrow Rx] \}$$

or, if preferred,

$$\forall x [Px \rightarrow (\neg Tx \leftrightarrow Rx)]$$

Sentences (40) and (41) can be deemed as bilateral reduction sentences for R, or definition (1).

Conclusions

In the literature, it has already been shown that it is possible to build reduction sentences for the definition of modulation the theory of mental models provides, that is, for (1). In particular, given that the characteristics that can modulate the models or possibilities of a sentence are three (content, context, and knowledge), three reduction sentences can be constructed, each of them capturing one of those features. They are sentences (22), (23), and (24).

The problem is that, as also described in the literature, those reduction sentences can be linked to sentences in natural language having one of the three characteristics and excluding the other two characteristics, that is, to sentences such as (14), (17), and (19). Thus, one can think about sentences modulated by virtue of content, and in which context and knowledge play no role, for example, (14). Likewise, there are sentences modulated by virtue of context, and on which content and knowledge have no influence, for example, (17). In the same way, it is possible to find sentences modulated by virtue of knowledge, and in which content and context do not act, for example, (19).

This is important because it seems to imply that bilateral reduction sentences are not possible for the definition of modulation in the theory of mental models. If each of the features that can cause modulation can exclude the other two features, biconditional relations cannot be established.

Nonetheless, a solution can be presented. If the three characteristics are negated at once and linked by means of conjunctions, as in (35) and (36), it is possible to build reduction pairs. Those reduction pairs can lead to a bilateral reduction sentence: (40), or (41).

It is not possible to construct bilateral reduction sentences for each of the characteristics that can modulate a sentence. This is because those characteristics, in certain cases, can imply that the other two characteristics are not present. But, assuming that the relation between the characteristics is disjunctive, that disjunction can be, as in (37), (38), and (39), taken as an only clause. This allows building a bilateral reduction sentence. Accordingly, the definition of modulation in the theory of mental models fits Carnap's (1936, 1937) framework even better than indicated in the literature.

References

- Byrne R M J, Johnson-Laird P N (2020) *If and or*: Real and counterfactual possibilities in their truth and probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 46(4): 760–780. <https://doi.org/10.1037/xlm0000756>
- Carnap R (1936) Testability and meaning. *Philosophy of Science* 3(4): 419–471. <https://doi.org/10.1086/286432>
- Carnap R (1937) Testability and meaning – Continued. *Philosophy of Science* 4(1): 1–40. <https://doi.org/10.1086/286443>
- Espino O, Byrne R M J, Johnson-Laird P N (2020) Possibilities and the parallel meanings of factual and counterfactual conditionals. *Memory & Cognition* 48: 1263–1280. <https://doi.org/10.3758/s13421-020-01040-6>
- Johnson-Laird P N (2010) Against logical form. *Psychologica Belgica*, 50(3/4): 193–221. <http://doi.org/10.5334/pb-50-3-4-193>
- Johnson-Laird P N (2012) Inference with mental models. In *The Oxford Handbook of Thinking and Reasoning*, eds K. J. Holyoak, R. G. Morrison, 134–145). New York, NY: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199734689.013.0009>
- Johnson-Laird P N, Khemlani S, Goodwin G P (2015) Logic, probability, and human reasoning. *Trends in Cognitive Sciences* 19(4): 201–214. <https://doi.org/10.1016/j.tics.2015.02.006>
- Johnson-Laird P N, Legrenzi P, Girotto V, Legrenzi M S, Caverni J P (1999) Naïve probability: A mental model theory of extensional reasoning. *Psychological Review* 106(1): 62–88. <https://doi.org/10.1037/0033-295x.106.1.62>
- Johnson-Laird P N, Ragni M (2019) Possibilities as the foundation of reasoning. *Cognition* 193. <https://doi.org/10.1016/j.cognition.2019.04.019>
- Khemlani S, Hinterecker T, Johnson-Laird P N (2017) The provenance of modal inference. In *Computational Foundations of Cognition*, eds G. Gunzelmann, A. Howes, T. Tenbrink, E. J. Davelaar, 663–668. Austin, TX: Cognitive Science Society.
- López-Astorga M (2022) The definition of modulation and its reduction sentences. *Open Insight*, XIII(27), 106–118. <https://doi.org/10.23924/oi.v12i27.506>
- López-Astorga M, Ragni M, Johnson-Laird P N (2022) The probability of conditionals: A review. *Psychonomic Bulletin & Review*, 29, 1–20. <https://doi.org/10.3758/s13423-021-01938-5>
- Popper K (2002) *The Logic of Scientific Discovery*. London, UK: Routledge. <https://doi.org/10.4324/9780203994627>
- Quelhas A C, Johnson-Laird P N (2017) The modulation of disjunctive assertions. *The Quarterly Journal of Experimental Psychology* 70(4), 703–717. <https://doi.org/10.1080/17470218.2016.1154079>
- Quelhas A C, Johnson-Laird P N, Juhos C (2010) The modulation of conditional assertions and its effects on reasoning. *The Quarterly Journal of Experimental Psychology* 63(9), 1716–1739. <https://doi.org/10.1080/17470210903536902>
- Quelhas A C, Rasga C, Johnson-Laird P N (2019) The analytic truth and falsity of disjunctions. *Cognitive Science* 43(9). <https://doi.org/10.1111/cogs.12739>

MORAL CHALLENGES FOR BAUER'S PROJECT OF A TWO-LEVEL UTILITARIAN AMA

Silviya Serafimova

Department of Ethical Studies,
Institute of Philosophy and Sociology,
Bulgarian Academy of Sciences,
Sofia, Bulgaria
email: silvija_serafimova@yahoo.com

Abstract:

The main objective of this paper is to demonstrate why AI researchers' attempts at developing projects of moral machines are a cause for concern regarding the way in which such machines can reach a certain level of morality. By comparing and contrasting Howard and Muntean's model of a virtuous Artificial Autonomous Moral Agent (AAMA) (2017) and Bauer's model of a two-level utilitarian Artificial Moral Agent (AMA) (2020), I draw the conclusion that both models raise, although in a different manner, some crucial issues. The latter are recognized as deriving from the complex relationships between human cognition and moral reasoning, as refracted through the lens of the idea of moral AI. In this context, special attention is paid to the complications which are triggered by the analogical thinking regarding the processes of replication of human morality in the field of machine ethics.

Key words: machine ethics, moral reasoning, Howard and Muntean's model of a virtuous AAMA, Bauer's model of a two-level utilitarian AMA.

1. Introduction

The main objective of my paper is to demonstrate why arguing for machines which effectively

perform moral tasks in the field of machine ethics does not necessarily mean that such machines can be coined moral machines by default. For the purposes of revealing the reasons behind this hypothesis, I investigate why human practical reasoning is irreducible to machine practical reasoning in moral terms. Specifically, my examination is focused upon comparing and contrasting Howard and Muntean's model of a multifaceted virtuous Artificial Autonomous Moral Agent (AAMA) (2017) and Bauer's two-level utilitarian project of an Artificial Moral Agent (AMA) (2020). Having previously analyzed the origin of some moral issues regarding Howard and Muntean's model (Serafimova, 2020), in this paper, I explore its reconsideration as refracted through the lens of Bauer's analysis.

1.1. Clarifications

1.1.1. Conceptual clarifications

Before going into detail in the comparative analysis of Howard and Muntean's and Bauer's projects, I would point out one key methodological specification which plays a crucial role for understanding what these two projects of moral machines look like. While Howard and Muntean's objective is to build a multifaceted virtuous AAMA, Bauer re-

sponds to their model with an investigation of a virtuous AMA, viz. he discusses a ‘weaker’ version of Howard and Muntean’s project. This specification raises the following methodological concern. Even if one agrees that Bauer’s model of a two-level utilitarian AMA provides more benefits than a virtuous AMA, it does not follow that such a model is *better* than a virtuous model of AAMA by default. The reason is that the latter displays a more elaborated version of a moral machine. In other words, Bauer’s statement that Howard and Muntean have developed a project of a virtuous AMA which they “prefer” to coin AAMA (Bauer, 2020, p. 265, Note 3) needs further clarification, since this is not a terminological specification having one and the same object of investigation.

If AAMA corresponds to what Moor coins *explicit ethical agent* — a machine that has the potential to calculate the best action in ethical dilemmas using ethical principles by itself — AMA meets Moor’s definition of *implicit ethical agent* — a machine that has been designed to behave ethically by following some ethical principles as embodied by its designer (Moor, 2006, pp. 19–20). The main moral difference is that while the AAMA is supposed to reach a human-level morality, the AMA is assumed to work with moral algorithms implied by its engineers. In addition to the concern that not every single algorithm related to morality is a moral algorithm in itself (Keeling, 2018), my main objection is that the two types of moral machines make room for different understandings of moral development. If an AAMA is supposed to achieve a human-level morality, this means that it should be able to reach a certain degree of *moral self-development* as well. In contrast, an AMA should reach a given level of *moral (self-)update* in the best possible scenario, since it has initially worked with preprogrammed algorithms related to morality.

Comparing and contrasting the development of the aforementioned two types of machines shows that even if a two-level utilitarian AMA is evaluated as *better* than a virtuous AMA in terms of providing better opportunities for *moral (self-)update*, it is initially incomparable with a ‘weaker’ version of an AAMA. This is due to the fact that the AAMA aims at reaching the point of morally developing itself, viz. achieving *moral self-development*.

However, taking into account that building AAMAs is still an impossible mission, the question is whether or not a *good moral (self-)update* can be

recognized as *better* than *imperfect moral self-development* in the field of machine ethics. If one refers to Moor’s clarification that explicit ethical agents act “*from* ethics”, not “*according to* ethics”, as do implicit ethical agents (Moor, 2009), the issue is whether or not working on resolutions when principles are in conflict is *better* as compared to working with preliminary embedded ethical considerations — specifically, when wrong resolutions are adopted in the end. Certainly, one could say that the choice of the embedded ethical considerations is safer, but it does not guarantee that the process of decision-making could develop over time when new, morally significant dilemmas occur. Furthermore, arguing for moral self-development raises the issue of why one would prefer to develop moral machines rather than strengthen interhuman moral development.¹

1.1.2. Introductory comments upon Howard and Muntean’s project

To provide context for my comparative analysis, I briefly outline what I consider to be some of the most important issues regarding Howard and Muntean’s project of a virtuous AAMA. Specifically, I maintain that their theory of why such a machine should not be designed as a global moral machine is a crucial contribution to the debates about moral AI. This means that Howard and Muntean’s project does not encourage one to decode human moral behavior in its complexity, but narrows the focus to the potential similarities between humans and machines.

In the language of Moor (2009), a virtuous AAMA can be described as a ‘weaker’ version of an explicit ethical agent. The strength of the ‘weakness’ can be exemplified by Howard and Muntean’s statement that the virtue ethics of AAMAs has a “partial analogy with human virtue ethics” (Howard and Muntean, 2017, p. 141). However, being unable to ‘read’ the complexity of human moral behavior, such a machine would not be able to properly ‘read’ and conceptualize human moral motivation or moral feelings. Furthermore, building a moral machine based on the computation of potential moral similarities between humans and machines triggers the

¹ For HA-AI trust relationships and the role of AI in building multi-agent systems, see Serafimova, 2021, p. 402, pp. 407–409.

question of whether or not there is a risk that this 'weaker' version of an explicit ethical agent may become as vulnerable as humans, who are imperfect moral agents.

Consequently, the methodological strength and weakness of Howard and Muntean's project can be defined as having one and the same crossing point, viz. what the authors refer to as a computational "*minima moralia*" for AAMAs (Ibid., p. 131). First, the methodological benefits concern the way in which moral self-update grounded in virtues contributes to minimizing the risks of blurring the categories of 'useful' and 'good' (as displayed by act and rule utilitarian approaches) in the process of behavior reading. This self-update also minimizes the restriction of moral agency to the update of duties alone (as can happen if a model of deontological AAMAs is adopted uncritically). Second, the specification that AAMAs' active learning should be grounded in a pattern recognition structure, which includes "non-linearity, emergence, errors, noise, or irrelevant data" (Ibid., p. 143), makes the project less ambitious in terms of moral perfection. Thus, it increases the possibility of updating and correcting some data if the latter is considered as unreliable in moral terms.

Regarding the methodological disadvantages, the main issue concerns the initial adoption of the so-called black-box strategy (Ibid., p. 131). This strategy affects the circumvention of the replicability of the mental and psychological mechanisms of moral judgments as well as the rational and deductive aspects of morality (Ibid.). However, limiting the analysis to the behavioral replicability of moral actions hides the risk that the analogy between humans and moral machines can be used as an excuse for machines' moral imperfection and associated moral imperfectability for the sake of grounding their autonomy.

In turn, the particular implications of the risks of having an immoral AAMA are traceable to the two main processes of computation regarding the behavior reading and systematization of data. Both processes initially depend upon the biases of the programmers and the so-called (by the authors) evolvers — people who make the decision about the evolutionary computation and who can be a different person than the programmer or the trainer of the neural network (Ibid., p. 152). Consequently, the evolvable features of each AAMA are associated by Howard and Muntean with the so-called robo-virtues

(Ibid.).² In this context, one of the main moral issues is that if robo-virtues can radically differ, they cannot guarantee that the scenarios of a virtuous AAMA will not turn into immoral scenarios in the process of the moral self-update.

2. Bauer's critical reception of Howard and Muntean's project

The main objective of Bauer's project is to demonstrate that a two-level utilitarian AMA "can account for nearly everything the virtuous AMA can account for, while also issuing some advantages" (Bauer, 2020, p. 267). Before specifying the gist of these advantages, Bauer discusses the three analogies employed by Howard and Muntean for their project of a virtuous moral machine. Specifically, he examines the implications of the analogies in question by assuming that "behavioral rules are akin to having dispositions associated with virtues" (Ibid., p. 265).

Adopting such an assumption, however, raises some methodological concerns about whether or not one can avoid moral relativism (moral particularism in Bauer's sense) when addressing machines' moral (self-)update. In this context, I try to prove that since there are many behavioral rules which do not necessarily represent the coupling of dispositions with virtues,³ arguing for the benefits of a two-level utili-

² Elaborating upon the debate about the building of virtuous moral machines assumes that one should shed light upon the so-called technomoral virtues (Vallor, 2016; Hagendorff, 2020, p. 112). According to Hagendorff, ethics can no longer be understood as a "deontologically-inspired tick-box exercise", but as a project advancing personalities, changing attitudes, strengthening responsibilities and most importantly, refraining from unethical behavior (Hagendorff, 2020, p. 112). Technomoral virtues as such include, but are not limited to, honesty, justice, empathy and care (Vallor, 2016). They are "supposed to raise the likelihood of ethical decision-making practices in organizations that develop and deploy AI applications" (Hagendorff, 2020, p. 113). Judging by the aforementioned clarifications, one may argue that technomoral virtues are *human* moral virtues, which are applied for the sake of grounding AI's use in a morally justifiable manner rather than being virtues in themselves.

³ For his analysis, Bauer relies upon the assumption of the correspondence between behavioral rules related to virtues and behavioral dispositions associated with virtues. He refers to the theories of Hooker (2000, p. 76, pp. 90–91) and Hare (1983, p. 39) (Bauer, 2020, p. 265). One of the main methodological concerns is that such a corre-

tarian AMA cannot be unquestionably accepted even within the framework of utilitarianism. An illuminative example in this respect is the well-known debate about the role of the dispositions to good and pleasant, which do not coincide by default.⁴ Extrapolating the discussion to the field of machine ethics, it would mean that the ‘content’ of the two-level utilitarian AMA triggers different consequences if behavioral rules are set within the so-called strong or weak rule utilitarianism.⁵

Bauer exemplifies Howard and Muntean’s virtuous moral machine model by a diagram (Ibid., p. 266). The diagram displays a “conceptual map” (Ibid.) which is based upon his interpretation of the prototype features of Howard and Muntean’s project. According to Bauer’s conceptual map, the virtuous AMA is provided with some initial input which concerns a training data set. Consequently, the set of data is used as a trigger for a machine

spondence may contribute to recognizing the moral implications of utilitarianism, but it is not unquestionably justifiable when applied to virtue ethics. Specifically, virtue consists of a behavioral disposition, viz. it is an excellent trait of a character, but such traits do not have the status of behavioral rules. For instance, an honest person cannot be identified simply as the one who is ‘disposed’ to telling the truth. In addition, they should also speak the truth and do the right thing. In turn, regardless of some *prima facie* similarities between utilitarianism and virtue ethics, one cannot argue for equating eudaemonist virtue ethics with hedonism nor can one identify the so-called target-centered virtue ethics with act utilitarianism. Roughly speaking, the disposition to virtues in virtue ethics is more agent-centered than within utilitarianism where the rules or actions are of primary importance.

⁴ The discussion about the mutual relatedness between good and pleasant has more apparent implications in the field of act utilitarianism; specifically, within hedonism. However, one does not have to neglect the fact that since rule utilitarianism is grounded in the expected maximization of happiness, the blurred boundary between good and pleasant may trigger some ambiguous consequences which do not necessarily have a purely moral origin.

⁵ According to strong rule utilitarianism, moral rules should be universally valid, which raises the risk that the two-level utilitarian moral machine can be grounded in a certain type of moral absolutism. If so, however, Bauer’s moral functionalism (Bauer, 2020, p. 267) loses its constructive normative potential. In turn, the risk deriving from weak rule utilitarianism is that the latter can deteriorate into act utilitarianism. Extrapolating the statement to Bauer’s model, this means that the two-level utilitarian AMA can be reduced to a one-level model, which is considered inappropriate by Bauer himself. For the reasons for choosing a two-level rather than one-level utilitarian AMA, see Bauer, 2020, p. 268.

learning process. The latter is based upon adopting neural nets+evolutionary computation (NN+EC) whose objective is to detect and evaluate patterns relevant to particular normative domains. In turn, the detection’s objective of machine learning is to achieve the recognition and the categorization of moral behavioral patterns. These patterns are representative of the justification of behavioral dispositions (of moral virtues) which should ‘guide’ the decision/action as an output. In brief, by utilizing multiple neural nets, the virtuous AMA is supposed to “perform evolutionary computational operations on their outputs, thus getting the best outcomes over an array of neural nets” (Ibid.).

In this context, the idea of virtuous AMAs’ moral (self-)update is interpreted by Bauer as a matter of gathering new data (which is processed by detecting new moral behavioral patterns) and categorizing its particular situation. The latter underlies the determination of the best decision or course of action (Ibid.). Thus, moral (self-)update is recognized as being dependent upon the continuous new input which brings the machine up to speed in new situations.

Discussing the three analogies offered by Howard and Muntean, Bauer provides the following specifications. The first analogy concerns the assumption that exercising a moral virtue is similar to exercising a practical skill (Ibid., p. 265). This analogy is generally supported by Bauer who adds that “building up a set of reliable conditional rules to follow” is also a matter of practical skill (Ibid.). That is why one needs to habituate oneself in the process of rule-following. Accepting such an interpretation, however, means that moral reasoning is similar to practical reasoning, which is quite a problematic stance.

If the exercise of a moral virtue is reduced to the habituation of rule-following, the idea of moral maturity will not assume a process of moral development, including the personal development of moral agents. In turn, rule-following as applied to virtues approximates the virtue ethical approach not to the utilitarian one, but to the deontological approach, as well as raising the risks of moral absolutism. Practically speaking, such an approximation entails the risk that one can generalize rule-following as a matter of duty, which does not necessarily require one’s moral convictions to be involved in the act of rule-following itself.

If moral (self-)update is based upon the aforementioned habituation, it would mean that the AMA should employ a rule as a virtue. Then, the moral (self-)update would be narrowed to an update of rules, making no room for understanding what the AMAs' dispositions should look like. In addition, one should also keep in mind that behavioral dispositions are not necessarily grounding moral virtues, since they may have a broader origin and wider impact due to the complexity of agents' general motivations.

Going back to Bauer's analysis, he accepts the second analogy provided by Howard and Muntean as well. This analogy is grounded in the assumption that machine learning is like human-skill acquisition because artificial learning and brain-based learning are "more similar than different" (Ibid.). However, the issue is that the process of collecting data for the purposes of discovering patterns is only a necessary condition for arguing for moral reasoning as such.

Furthermore, if we agree with Bauer that "many cognitive skills can be captured as sets of conditional rules" (Ibid.), this would mean that first, both humans' and machines' moral motivations are simplified, since moral feelings and biases cannot be classified by adopting conditional rules. Second, the design of moral machines would be focused upon their autonomy rather than morality due to the fact that conditional rules can guarantee a certain level of autonomy by analogy with that of humans. In other words, reducing human morality to machine morality on the basis of a cognitive scheme of rule-following can lead to grounding a limited moral epistemology. The latter does not necessarily favor AMA's moral (self-)update because it cannot encompass the complexity of human moral motivation either.

Probably one of the most questionable exemplifications of *human-moral machine* replication can be found in the so-called moral biomimetics (Ibid.). This replication is recognized as rooted in the assumption that since there is no special human faculty for moral cognition, "machines can reproduce morally virtuous behavior" (Ibid.). Even if one neglects the difficulties in replicating feelings, emotions and biases into machines, the issue remains as to how these machines can avoid human moral imperfections and consequently, whether or not it is possible to argue for moral machines' moral (self-)update at all if biomimetics is the objective of their development.

Bauer's own interpretation raises some new questions as well. He argues that if "mimicry" [biomimetics] is the goal, moral cognition can also be displayed by the sets of conditional rules and the recognition of the circumstances in which they apply (Ibid., p. 266). Following such a scenario, however, would mean that the AMAs can easily fall into the trap of moral relativism (if the circumstances are considered as a leading factor for decision-making) or into that of moral absolutism (if the role of conditional rules is exaggerated).

The origin of some of the aforementioned concerns can be found in what Bauer defines as a background of metaphysical and metaethical assumptions of Howard and Muntean's model, viz. moral functionalism and internal quietism (Ibid.). Bauer himself raises this discussion in relation to Howard and Muntean's third analogy, saying that roughly equating moral cognition with general cognition is possible if moral thinking is determined merely to be "thinking applied to moral questions" (Ibid.). In this sense, the choice of moral functionalism is grounded in the striving for achieving objectivity within the input-output process regarding moral issues. However, guaranteeing the objectivity of moral questions does not mean that such an objectivity is 'moral' by default.

Howard and Muntean themselves choose moral functionalism on the grounds that it favors the supervening of moral properties on natural properties (Ibid.). Accepting such a correspondence is used for the purposes of giving normative grounds for the initial input (the training set of data). Consequently, the change of moral properties is justified as necessitated by a change of natural properties. This change should guarantee the normative validity of the (self-)update of virtuous machines as a matter of a continuous new input. Moral functionalism, however, does not explain why humans face the pitfalls of moral pluralism regarding one and the same natural properties.

The apparent challenges of moral functionalism in building virtuous AMAs can be better traced to the so-called internal quietism which concerns the gist of "the functional equivalence of cognition" (Howard and Muntean, 2017, p. 134; Bauer, 2020, p. 266). Elaborating upon such a formal criterion of equivalence would mean to misleadingly equate the role of moral reasoning with that of moral cognition. On the other hand, even if one argues for a specific concept of moral cognition, it does not guarantee

that moral reasoning will be applicable to moral cognition by default. This is due to the fact that moral reasoning is closely tied with the role of complex moral motivation, which is not purely cognitive by default either. Specifically, moral reasoning may rest upon different types of biases⁶ which are not necessarily of a purely moral origin.

Paradoxically enough, moral absolutism may be driven by the assumption that every single situation requires one to adopt different mechanisms of detection and evaluation in moral terms, which are recognized as being universally valid for the particular moral machine. This means that one should operate with different moral machines in different situations. Arguing in theoretical terms, the origin of the risks can be traced back to the invisible transformation of moral particularism into moral absolutism. The transformation itself results from the striving for moral generalism, which is supposed to introduce ‘more’ objectivity into the field of machine ethics.

3. Bauer’s project of a two-level utilitarian AMA

According to Bauer, the two-level model of AMA is supposed to combine the virtues of rule utilitarianism with those of act utilitarianism. While the implications of rule utilitarianism (Rule U) can be found on the first level (level 1), viz. when moral decision-making requires the following of a set of rules, act utilitarianism (Act U) addresses the challenges on the second level (level 2) — that of the application of critical thinking when the rules conflict or do not exist in given cases (Bauer, 2020, p. 267).

As discussed in the previous section, Bauer accepts Howard and Muntean’s assumptions that (1) moral cognition is “sufficiently analogous” to general cognition (their third analogy) (Ibid.), and that (2) moral functionalism and internal quietism are reasonable metaphysical and metaethical assumptions. In addition, he also specifies that the other two analogies formulated by Howard and Muntean “are not immediately present” in the two-level utilitarian AMA (Ibid.). Bauer does not deny the analogy between machine learning and human-skill acquisition (the second analogy), but argues that it is not neces-

sary. In turn, this does not mean that “a moral virtue is not like exercising a practical skill” (Howard and Muntean’s first analogy). However, such an analogy is not implemented, since the two-level model does not require virtues either (Ibid., pp. 267-268).

If the two-level utilitarian model is built within the framework of moral functionalism, which is supported by the analogy between moral and general cognition, one cannot understand why the analogy between machine-learning and human-skill acquisition is not necessary. In other words, it remains unclear how one can figure out when and why the analogy regarding learning and skills is not necessary, taking into account that one has already adopted the general knowledge analogy as such. Furthermore, if the two-level model does not require virtues and consequently, does not tackle the reasons behind Howard and Muntean’s first analogy, nothing guarantees that one can provide rule utilitarianism with a “virtue-like normative flavor” (Ibid.: p. 268).

Comparing and contrasting Bauer’s conceptual map of Howard and Muntean’s project and his own shows that the difference concerns the evaluative model, which entails level one (rule utilitarianism regarding the rule $X \rightarrow Y$ (X: conditions satisfied, Y: moral decision/action)) and level two (critical considerations when there are contradicting rules or a lack of rules (X not satisfied, so calculate utility and form a new rule)) (Ibid.). Moral virtues are not included on the level of behavioral dispositions, as mentioned above, since they are replaced with conditional rules (Ibid.).

Regarding the first level, the two-level utilitarian AMA should gather data from “its ‘moral situation’ or environment” (Ibid.). “If an established rule fits the situation, the rule is triggered by satisfaction of the required antecedent conditions”: applying if X (some possible set of moral properties in the AMA’s environment), then Y (a decision or action as a response to the situation) (Ibid.). Before analyzing why the application of rule-utilitarianism is not necessarily analogous to the activation of a behavioral disposition in the virtuous AMA, I will try to shed light upon the specificities related to choosing sets of rules.

According to Bauer, rule utilitarian theorists are “generally interested” in defining “the definitively correct set of rules”, just as Kantians define the ‘correct’ moral duties and “virtue theorists aim to identify the ‘correct’ virtues” (Ibid.). However, the categories of rules, duties and virtues are not of a similar

⁶ For the role of biases in the field of machine ethics, see Serafimova, 2020a, pp. 114-115.

rank within moral theory. While rules primarily address the functioning of moral practice by regulating collectively determined moral agency, duties and virtues are mainly concerned with the particular behavior of the moral agents themselves. That is why the development of duties and virtues requires the development of individual and/or collective moral maturity. Practically speaking, rules are not duties nor are they necessarily based upon virtues, although in following one's duty, one can establish some rules regarding moral conduct. Consequently, one may witness many immoral rules which originate from moral virtues or moral virtues which are developed in an attempt to eradicate immoral rules. In turn, a duty may be fulfilled which is grounded in a moral virtue, but so can the one which may be grounded in a vice.

Arguing for epistemic clarifications in this context does not solve the aforementioned complications, since what is right in epistemic (economic, political etc.) terms is not necessarily good in moral terms. Specifically, most complications are due to the fact that some formally correct rules may result from morally incorrect duties or pseudo-moral virtues. That is why if one neglects the complexity of moral motivation and biases, one takes the risk of mistaking the categories of rules, duties and virtues as being of a similar rank.

Another methodological difficulty derives from Bauer's suggestion of "a workable approach" to getting AMAs "going in a morally fit direction", taking into account that AMAs' actions work relevantly only in respect to "certain populations of individuals" (Ibid.). In this context, the main issue is how in respecting the principles of cultural diversity one can avoid moral relativism (particularism) when arguing for moral machines. For instance, if one team of experts provides guidelines for such AMAs respecting the values of a given culture, does this mean that such machines can work as moral machines only within 'their' own culture, considering that even the rules within one and the same culture vary (e.g. over time)?

Regarding the internal limitations of culture, one may argue that the risk of moral particularism gains new strength. Extrapolating the debate to the two-level AMAs shows that the role of the second level, that of critical thinking, is initially limited, since critical thinking should also be non-contradictorily defined in cultural terms. In other words, the calculation of expected utility when the

initial rules contradict or are unusable should be primarily determined by the given culture. If the particular cultural background is changed, this would make such moral machines inapplicable.

In addition, the moral responsibility of those who are supposed to design the AMAs' initial rules should be analyzed as well. Addressing this issue brings us back to the previously discussed role of engineers' biases. According to Bauer, experts "in relevant fields, combined with feedback from relevant affected parties" (businesses, governments, individual citizens etc.), can provide some guidelines (Ibid.). However, leaving such a collaboration unspecified in moral, economic, cultural and socio-political terms introduces the risk that rules can be chosen for the sake of prioritizing some political, economic and governmental interests which are not necessarily moral. One should keep in mind who takes responsibility when the choice of rules is driven by conflicting interests, (viz. who decides that those rules should be taken as initial rules for moral behavior) and observe how various stakeholders and groups of people take sides in such debates.

Going back to the analysis of the categories of rules, duties and virtues, one may argue that the problem of expertise can lead to subjecting duties and virtues to rules and thus undermine their normative validity. In turn, adopting such an approach would negatively affect the design of the two-level AMAs, viz. they will be utilitarian rather than moral.

Regarding the role of the second level, viz. that of critical thinking concerning the application of the utility calculus, one may witness a problem pertaining to the effects upon the addressees. According to Bauer, the utility calculation "must incorporate the effects of an action on all relevant individuals" (Ibid.). Conflicts between rules should not be solved by applying act utilitarianism, as Hooker argues (2000, p. 89), although the reason is not, as some might assume, that people would lose confidence in the system of rules when arguing for machine ethics (Bauer, 2020, p. 268, Note 8). On the other hand, regardless of the fact that confidence is not an issue for the machines (Ibid.), it does not follow that the issues related to the introduction of level 2 should be neglected.

The concerns about moral particularism are raised again but this time due to the problems with the moral machines' (self-)update. Certainly, humans do not know all the correct rules that the AMA will need (Ibid., p. 269). That is why the crux of the

problem lies in how the introduction of level 2 will help the AMA to adapt to novel situations (Ibid.). If humans do not know all the rules that govern moral behavior in all possible situations, how can we expect machines to be able to attain better knowledge – keeping in mind, moreover, that moral knowledge is not sufficiently analogous to cognitive knowledge? Specifically, what would guarantee that such machines would be able to predict the complexity of the effects of morally salient acts to all the relevant addressees, taking into account that humans are also unable to do this? In other words, if humans as full ethical agents (Moor, 2009) lack this ability, what hope is there that machines will achieve it in terms of analogical moral thinking?

Another equally significant task is proving that the two-level utilitarian machines provide more benefits than the virtuous ones. Some of the main concerns about the analogy between ‘correct’ rules, duties and virtues in moral terms have already been examined. In addition, I would also argue that the analogy between virtues and rules obtains when one accepts a too broad definition of psychological disposition as a moral disposition.

As Bauer cogently demonstrates, one should avoid discussions of AMAs’ feelings (Bauer, 2020, p. 269). Adopting such a line of thought approaches a certain type of constructivist formalism according to which “it is to behave as if one had the relevant feelings by following rules” (Ibid.). However, this model displays a distorted and limited picture which makes one unable to see the complexity of psychological dispositions as a main premise of the complexity of moral motivation.

Bauer himself seeks to illustrate the connection between moral virtue as a matter of psychological disposition and having a certain psychological disposition as being associated with the conditional behavior of following a conditional rule (Ibid.). Analyzing his theory of how having a moral virtue implies the following of a rule, I would argue that moral concerns and consequently, those about the two-level utilitarian AMAs, derive from the understanding of dispositions in epistemological terms, as well as recognizing the corresponding similarities to some rules on an epistemological basis alone.

Bauer introduces a table in order to demonstrate how dispositions and rules have a similar logical structure (Ibid.). Certainly, such a structure can be found in the process of formalization, when one distinguishes between a triggering condition, dispo-

sition and manifestation of this disposition (Ibid.). However, the example he chooses — a hammer striking a fragile object and, consequently, the object breaking — is dissimilar to the normative example to which it is compared — a question being asked of an agent who is honest (a virtue) and the agent, consequently, responding truthfully (Ibid.). I would argue that the example of physical causation (i.e. if an object that is fragile is hit by a hammer, the object will break) is more likely to hold than the moral example (i.e. if an honest person is asked a question, the person will “tell the truth” (Ibid.).

Why is this so? One of the reasons is that virtue is not equivalent to a rule in normative terms because it is a character trait., not a natural law, and it influences but does not determine behavior. For the triggering condition (a question being asked) to produce a truthful response, a person must first be convinced (morally, socially, politically etc.) that if a question is asked, they should tell the truth, and even when this is the case, other factors may constrain or influence the agent’s behavior. In contrast, if the object is fragile, it simply gets broken when it is hit hard enough due to the laws of physics.

If one follows Bauer’s theory of the logical similarities between dispositions and rules, would it be possible for an AMA to activate level 2 if a white lie is needed? The crux of the question aims to address how the calculation of good can be not only conducted, but also properly evaluated in this case. If utility is calculated as ‘pleasure’, this would mean that an AMA should lie in order to hide the painful truth. However, this runs the risk of the white lie (which aims to reduce harm) devolving into mere deception (which causes harm). On the other hand, if utility is considered as ‘good’, then telling the truth, which may cause pain, may not be the morally right choice.

Would it be possible for an AMA to develop a moral algorithm that determines when telling the truth is the right choice, taking into account that the difference between a white lie and mere deception depends upon the agent’s moral motivation? That is, would such an AMA manage to update itself in time and space within a given culture and make moral decisions informed by the assumption that all parties might have feelings which are not ‘included’ in the process of computation? If so, is it possible that the AMA’s moral (self-)update would lead to a paralyzed inability to make moral choices due to the fact

that the individuals it affects are not a homogeneous group of moral agents either?

Even if rules become part of the AMAs' 'moral' character, it does not follow that these machines will become moral, since, as was shown above, the relationship between autonomy and morality is quite complicated in the field of machine ethics. In turn, Bauer's conception of pre-loading machines with "some of the most widely agreed upon general rules" (viz. with our moral rules, not with our behaviors), as a trigger for making them "potentially... better than typical human behavior" (Ibid.) requires further clarification.

Such an assumption is correct if one assumes that immoral behavior results only from not following rules or breaking rules. However, these two processes do not have an equal normative validity. It is one thing to break rules and another to not follow them without violating them. On the other hand, both cases may lead to the moral reconsideration of outdated moral rules which do not reflect the current cultural, socio-political, economic and moral situation.

If virtuous AMAs are imperfect because they are intended to match human behavior, as Bauer argues, moral perfection is assumed to derive from the way in which the two-level utilitarian AMAs follow ethical rules. These machines, then, do not differ much from the deontological machines, since they adopt a similar line of constructivist functionalism, including the duty to follow rules. However, the risks of building immoral two-level utilitarian machines are higher in the process of moral (self-)update. This is due to the second level of act utilitarianism, which can make room for immoral calculations that rest, for instance, on the intention of identifying good with pleasant.

In addition, if having a moral virtue (which consists of a relevant behavioral disposition) is similar to following a rule, it remains unclear why Bauer is not concerned when saying that any two-level utilitarian AMA avoids the explicit references to virtues due to the assumption that virtue is "a relatively intangible concept that only complicates the conceptual design of the AMA" (Ibid.). At the same time, he considers as an advantage the assumption that the two-level utilitarian machine — specifically, its level of rule-following — can approximate the dispositions of the virtuous AMA (Ibid.). In this context, it also remains unclear what the particular benefits of the two-level utilitarian AMAs are. If rule

utilitarianism (level 1) approximates a virtue-ethical approach, then the unique benefits should be sought on the second level (level 2), viz. that of critical thinking, which is the level of act utilitarianism. The investigation shows that all these concerns do not give sufficient grounds for one to favor act utilitarianism over virtue ethics in the field of machine ethics.

Judging by the aforementioned examinations, I would argue that the main difficulty in justifying the moral (self-)update of a two-level utilitarian AMA is that the update in question is evaluated through the argument for its imperfection by analogy with human imperfection. One of the crucial issues regarding the argument for imperfection is that it relies upon a vague probability of future improvement. Consequently, the major concern about the latter clarification is that human mistakes and imperfections can be tacitly recognized as epistemological and moral excuses for machines' errors.

Specifically, the complicated relationship between human mistakes and computational errors is unsolvable if one focuses upon improving merely the errors. Bauer himself considers the crux of the mistakes mainly as arising from technological challenges. According to him, one can see an ethical issue for the two-level utilitarian AMA in its ability to identify and classify "morally relevant features in its environment" (Ibid.). However, such a specification is just the tip of the iceberg. What one can say further is that ethical concerns are deeply connected with the responsibility of AMAs' designers as moral agents and only then, with their competence as experts in computing the aforementioned morally relevant features.

Bauer argues that computer programming "will very likely continue to improve"; machines "will continue to get better at making very fine-grained, rapid distinctions regarding input from their environment, and they will be better able to account for suitable contextual factors" (Ibid., p. 270). This technological optimism is grounded in the idea of progress whose prototypical feature is technological advancement. However, one should keep in mind that the development of moral machines at that stage shows that humans are still quite far from granting moral autonomy to machines.

Bauer continues that "it is worth noting that humans do not always make moral distinctions and discern morally relevant features very well" (Ibid.). Certainly, the distinction between friend and foe on

a modern-day battlefield (Ibid.) depends much upon individual psychological status and environmental conditions. On the other hand, shifting the debate to the field of machine ethics does not mean that one can ground arguments for moral development in the analogy of human and machines' moral imperfection. Adopting such an approach requires one to assign moral responsibility in more subtle terms. This means that humans are morally responsible not only for discerning between friend and foe, but also for clarifying how machines can do so. The idea that the two-level utilitarian machine is not perfect, but no model of AMA is (Ibid.), does not shed light upon the issue of *when* and consequently, *how* one can stop such a machine if it becomes unacceptably imperfect in moral terms.

Extrapolating the debate to a meta-methodological level, the issue is *when* and *how* machines' moral imperfection should be limited and if so, why do we need to design such machines which can do nothing better than humans in the field of morality? Is there a risk that the technological optimism about machines' moral imperfectability encourages the building (either deliberately or inadvertently) of a future dystopian society? Moreover, who should take responsibility for the possible negative outcomes? Following such a line of thought brings us back to the risks of building hyper-realities in Baudrillard's sense in which the simulacra of ethical deliberation expose an existential threat to both human autonomy and human morality (Serafimova, 2020).

4. Conclusion

Of the many moral concerns that are raised by Bauer's analysis of Howard and Muntean's model of a virtuous AAMA as well as Bauer's own hybrid model of a two-level utilitarian AMA, I have focused on issues regarding the similarity between the ways in which behavioral rules and behavior dispositions are related to virtues. Elaborating upon Bauer's line of thought, I would argue that such a similarity holds only within the field of the questionable moral functionalism. The main concern here is that moral functionalism encourages the operation of moral formalism, which is based upon a limited kind of analogical thinking connecting moral cognition and general cognition.

In this context, debates about the role of moral feelings and moral motivation as well as those about the competence of the engineers who design moral machines, face the following challenges. Regardless of the fact that Bauer intentionally avoids the issue of moral feelings in his project, the acceptance of Howard and Muntean's theory of moral biomimetics demonstrates why such mimetics is one of the most troubling issues within moral formalism.

In general terms, the implications of moral formalism affect the initial recognition of the epistemological (and consequently, that of the moral) similarity between the categories of duties, rules and virtues, which are defined as being of a similar rank. However, one should keep in mind that while rules primarily address moral practice as a collectively determined moral agency, duties and virtues, by contrast, are related to the particular behavior of moral agents. Specifically, duties and virtues affect the development of agents' moral maturity. Practically speaking, rules are not duties nor are they necessarily based upon virtues, although in following one's duty, one can establish some rules for moral conduct.

Judging by the aforementioned investigations, I draw the conclusion that the difficulties in examining the similarities between behavioral rules and behavioral dispositions on the one hand and virtues on the other are due to Bauer's assumption that these similarities will be recognized as epistemologically related. However, epistemologically grounded similarities do not display moral similarities by default. A proof of this statement is found in the justification of Bauer's own model.

Analyzing his model, I would argue that there is a significant moral difference on the level of rule utilitarianism (the first level of Bauer's model) if the adoption of the formula $X \rightarrow Y$ concerns the fragility of an object that is hit with a hammer or an agent telling the truth when asked a question. Certainly, the laws of physics are not equivalent to those of morality. The object will be smashed if the hit is hard enough and this is a sufficient, objectively determined condition. In contrast, in the case of telling the truth, the moral agent should be convinced in one way or another that they *should* tell the truth. However, telling the truth is not a sufficient condition for determining whether the agent has acted in a moral way since they may be guided by some immoral or amoral motivation, such as avoiding punishment, being appreciated by others etc.

Applying these concerns to the field of machine ethics does not make the case less complicated either. The 'correction' of the insufficiency of rule utilitarianism (level 1) by act utilitarianism (level 2) raises the risk of moral particularism. Paradoxically enough, Bauer's idea of avoiding moral relativism by reconsidering the role of moral generalism for the two-level utilitarian model of AMA causes the concerns about the risk of relativism to grow. In addition, one should also keep in mind that the origin of such risks can be found on the first level of rule utilitarianism, when Bauer argues that the rules fed into the AMA might even differ between different cultures. Due to the fact that cultures themselves are not static phenomena, but develop in time and space, the issue is that one may build two-level utilitarian AMAs which are valid only for given cultures and consequently, only for a given period of time.

Elaborating upon the issues regarding the *first level* of rule utilitarianism (the implications of moral formalism in terms of behavioral rules and behavioral dispositions, as related to virtues) and the *second level* of critical thinking (specifically, the risks of moral relativism, which result from machines' lack of moral motivation and moral feelings) necessitates the examination of the two-level utilitarian AMA's moral (self-)update. Exploring the reasons behind Bauer's statement that computer programming will very likely continue to improve, and given the belief that humans cannot always make choices that have morally perfect outcomes either, I would claim that the main difficulty in justifying such a moral (self-)update is that it is evaluated through the lens of the argument for imperfection. In turn, the major issue regarding this specification is that human mistakes and imperfections are tacitly recognized as epistemological and moral excuses for machine errors.

Arguing for correctness, however, does not resolve the aforementioned complications since what is right in epistemic (economic, political etc.) terms is not necessarily good in moral terms. This means that most complications are due to the fact that some formally correct rules may result from morally incorrect duties or pseudo-moral virtues. That is why if one neglects the complexity of moral motivation and biases, one runs the risk of defining the categories of rules, duties and virtues as being of an equal rank.

In conclusion, comparing and contrasting Howard and Muntean's and Bauer's models of moral machines shows that human cognition does not nec-

essarily coincide with moral reasoning. Consequently, cognition regarding moral reasoning does not necessarily represent the moral specificities of such a reasoning, nor is every practical reasoning moral reasoning in itself. Furthermore, the encouragement of analogical thinking, which grounds the idea of replicating human and machine morality, faces the serious concern about the lack of relevant analogies between humans and machines in moral terms.

References

- Bauer W A (2020) Virtuous vs. Utilitarian Artificial Moral Agents. *AI & Society* 35: 263–271.
- Hagendorff T (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines* 30: 99–120.
- Hare R M (1983). *Moral Thinking: Its Levels, Method, and Point*. Oxford: Oxford University Press.
- Hooker B (2000) *Ideal Code, Real World*. Oxford: Oxford University Press.
- Howard D, Muntean I (2017) Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency In *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics*, ed. T. M. Powers, 121–159. Springer.
- Keeling G (2018) Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles In *PT-AI 2017, SAPERE 44. Studies in Applied Philosophy, Epistemology and Rational Ethics*, ed. V. Müller. Cham: Springer. https://doi.org/10.1007/978-3-319-96448-5_29.
- Moor J H (2006) The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21(4): 18–21.
- Moor J H (2009) Four Kinds of Ethical Robots. *Philosophy Now. A Magazine of Ideas* 72 https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots.
- Serafimova S (2020) Whose Morality? Which Rationality? Challenging Artificial Intelligence as a Remedy for the Lack of Moral Enhancement. *Humanit Soc Sci Commun* 7 (119). <https://doi.org/10.1057/s41599-020-00614-8>.
- Serafimova S (2020a) How to Assign Weights to Values? Some Challenges in Justifying Artificial Explicit Ethical Agents. *Balkan Journal of Philosophy* 12(2): 111–118.

Serafimova S (2021) Questioning the Role of Moral AI as an Adviser within the Framework of Trustworthiness Ethics. *Философия [Philosophy]* 30(4): 402–412.

Vallor S (2016) *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford: Oxford University Press.

MENTAL REPRESENTATIONS OF POLITICAL DISCOURSE IN AN AUTHORITARIAN SOCIETY

The case of Albania during the implementation of the Chinese Cultural Revolutionary model

Majlinda Bregasi, Albert Bikaj***

*University of Pristina "Hasan Prishtina"
Rr. "George Bush", Nr.31, 10000 Prishtine, Republika e Kosovës
e-mail: majlinda.bregasi@uni-pr.edu

**Faculty of Humanities, University of Zagreb
Jordanovac 110, HR-10000 Zagreb, Croatia
Website: <http://www.ffrz.unizg.hr/>
e-mail: bikaj@live.com

Abstract:

After the Second World War Albania was left under the Eastern Bloc. In 1967 Enver Hoxha, the leader of Albania from 1944 until his death in 1985, decided to implement the Chinese Cultural Revolutionary model. This article analyzes his speech, on February 6, 1967, before his comrades, who were supposed to be his eyes, ears, and mouth. It was in this way that his face, his thoughts and his words became ubiquitous throughout the country. In a highly authoritarian society political discourse has a direct and pervasive impact on peoples' lives and all aspects of the society.

This article is organized by analytical themes based on aspects of discourse analysis, but we have also applied the cognitive approach and imagology as auxiliary theories in order to achieve a better understanding of mental representations, especially the ones used to reinforce stereotypes about rural people. Given that these mental representations are still fostered by current politicians in order to establish power, we note how important it still is to analyze them. Considering that political discourse is a product of individual and

collective mental processes it is important to show, especially to younger generations, where these mental schemas come from.

Key words: discourse analysis; political speech; legitimacy; mental schemas; cognitive metaphors.

Introduction

After the Second World War Albania was left under the Eastern Bloc. In 1967 Enver Hoxha, the leader of Albania from 1944 until his death in 1985, after cutting his ties with Moscow decided to implement the Chinese Cultural Revolutionary model. This article analyzes a speech he made on February 6, 1967 to his comrades, who were supposed to be his eyes, ears, and mouth. In doing so, his face, his thoughts and his words became ubiquitous throughout the country. Foucault (1972) sees discourse as a form of constituting and constructing society; there-

fore, in a highly authoritarian society, like Albania in this case, political discourse has a direct and pervasive impact on peoples' lives and all aspects of the society.

The focus of this article is the examination of possible mental representations stimulated by this speech and how they are built up in order to establish legitimacy, authority, and political power and to cultivate uniform thinking. This article is organized by analytical themes based on aspects of discourse analysis, but we have also applied the cognitive approach and imagology as auxiliary theories in order to achieve a better understanding of mental schemas, especially the ones used to reinforce stereotypes about rural people.

Given that these mental representations are still fostered by current politicians in order to establish power, we note how important it still is to analyze them. Discourse, as a manner of saying and doing things, has become an indispensable instrument for political elites to manage a fluid but still highly dynamic and yet fragile society (Shirk, 2008, 2011). Considering that political discourse is a product of individual and collective mental processes it is important to show, especially to younger generations, where these schemas come from.

Truth production

Since Aristotle the language of politics has been seen as the rhetorical discourse par excellence. For Aristotle, rhetoric represented the ability to find in each argument features that have the capacity to create conviction. Rhetorical structures have a communicative function. Basically, they can be defined in terms of specific transformations of grammatical structure, such as additions, deletions, permutations, or substitutions, as in the case of alliterations, rhymes, or parallelisms at the morphosyntactic level, and metaphors, irony, or understatements at the semantic level (van Dijk, 1987: 35).

There is no doubt that if we are concerned with language and politics, we have to be concerned with truth and falsity in relation to a real world in which human interests and human suffering are real (Chilton, 2004: 49). According to Habermas, knowledge is not a neutral representation of an objective world, but it is the language that represents it. We can see linguistic behaviour as a medium through which rationality is achieved. Thus, truth is to be seen as a product of complex cultural, historical and psychological circumstances, and "to be understood as a system of ordered procedures for the production,

regulation, distribution, circulation and operation of statements" (Foucault in Rabinow, 1984: 74). Put briefly, meanings are in the minds of people and language-in-use is crucial.

The aim of political elites is to construct and present the political truth as historical truth in order to establish legitimacy, authority, and political power. The political elite in Albania redefined historical and cultural events in accordance with the imperatives of that period. According to Fairclough (1989:37), the role of a transformative discourse in a society that experiences radical changes is vital for sustaining those changes and especially for building uniform thinking. Thus, meanings are established as historical truth by discursive renaming and repositioning in relation to the progressive role of political elites in moving society forward (Qin Cao, 2014: 8). When humans interact, they do not communicate only with respect to social roles, boundaries, and bonds, but they also communicate representations of the world.

The objective of a political discourse is to promote the acceptance of an opinion through argumentation, or to generate acceptance of an opinion without argumentation to reach further conclusions. In the second case we are dealing with classical propaganda (Baldi, Savoia, 2010: 181). In this case the question is not whether the truth was spoken or not, but whether the public was convinced or not. Individuals are matching logical forms derived interpretively from the utterances produced by others to their mental representation of reality derived via perception and limited or coloured by their cognitive apparatus (Chilton, 2004: 50). The aim of language-in-use in politics is to modify this mental representation of reality.

Categories of strategic function

Let us look closely at this speech:

How can basic communist cells be in a revolutionary position, up in the mountainous village of Zgjan, the Hysgjokaj locality, in the district of Lushnja, when almost all the girls there, aged between 13-14 years old, are engaged and even married, illegally, by their parents? I will go even further. How can the communist cell of the Lezha district be in a revolutionary position, when the vile habit of selling girls for marriage has tacitly become common? (E. Hoxha, 1967: 3-4).

If we start off with the two rhetorical questions in the text we must deal with different ‘categories of the strategic function of language’, which Chilton and Schäffner (1997: 211) analyze as representation/misrepresentation and legitimization/delegitimization. Choosing the marriage of young girls as a topic of focus in a political speech, the political leader creates an image schema in which the Party represents itself as the only legitimate protector whose immediate duty is to save their lives.

Taking into consideration that image schemas are abstract patterns arising from a range of perceptual experiences and as such are not available for conscious introspection (Lakoff, 1987, 1990, 1993) and Johnson, 1987), they play a strong role in creating abstract thoughts. This is especially true in the case of Hoxha’s speech, where the data transmitted by the text was not verifiable, but was nonetheless provided as indisputable truth. The selection of words used to describe this location – ‘the mountainous village’ – is a misrepresentation since the district in question is located in a valley in the southern part of the country, and the name of the district is given just to reinforce it. The reason for using ‘the mountainous village’ is that it aligned with the lengthy propaganda campaign carried out against the northern part of the country, a mostly mountainous region. After the installation of Hoxha’s regime, the people of this region, most of whom were Catholic, were the biggest opponents of his regime. Thus, a strong propaganda campaign was fabricated against them. As the Catholic clergy took a firm anti-communist stance and never accepted Hoxha’s regime as legitimate, they were declared ‘enemies of the people’ on February, 11, 1945, and a majority of the Catholic elite were sentenced to death. As image schemas are multi-modal, schematic, and emerge from ongoing embodied experience (Lakoff, 1987, 1990, 1993) and Johnson, 1987), ‘the mountainous village’ brings to mind ‘the old enemies of people’.

The concept and legitimacy of the face

This speech, which was the first Hoxha gave in 1967, and which had a crucial influence on the changes planned for the country, begins with the demand that members of the Communist Party undertake the project of revolutionizing their lives. Hoxha addressed them in singular form, establishing a direct link with each communist.

Self-critique, as well as elections, are considered to be of great importance for the further

strengthening of the Party. For this purpose, in the meetings, ‘every communist should examine their work and that of the organization, measuring successes and failures in the spirit of criticism and self-criticism, set the objectives and fight for their accomplishments in the future ...’ (Hoxha, 1967).

As it is clear from the excerpt above, the speech starts with the imperative of self-criticism for all comrades, which is repeated several times during the text as a strategy of reinforcement. Considering Goffman’s face concept, we can see this imperative as a strategy for making communists lose their own positive value and for subtly constructing political legitimacy. For rulers, legitimacy means implementing their will and making rules and decisions acceptable, and they uphold legitimacy as a means of maintaining their status, prestige and boundaries (Gilley, 2005; Johnson, 1972; Macdonald, 1995).

Thus, Hoxha had to provide a strong emotional framework for his narrative by bringing together a ‘losing face’ with the image schemas of families that are selling little girls for marriage. This is a case where morals and politics come together perfectly to create intrinsically strong emotions and powerful effects of empathy. Since “empathy is morally powerful, and its political power seems to arise from its moral force, which in turn, is a consequence of brain structure [...] Morality is about right behaviour, behaviour that leads to well-being” (Lakoff, 2008: 103). As such, asking his comrades to carry out the moral duty of self-criticism is further justified by the possibility of ‘losing face’ – of having one’s social value undermined, usually in front of others who were eager to denounce and criticize, as good communists do. At the heart of nurturing and authoritarian models of families lie two central metaphors for morality: Morality as Care and Morality as Obedience to Authority (Lakoff, 2008: 105). The values of purity, care and obedience are dominant in this speech in order to strengthen the exercise of authority and reinforce a notion of morality as obedience to authority.

Cognitive approach

The cognitive approach considers political discourse as a product of individual and collective mental processes (Chilton, 2004: 51). In this regard, Chilton (2004) examines the roles of long-term memory (personal or social), which he also calls a recollection frame, and short-term memory (which deals with online processes) in serving as a storage for the contents of political discourse. Thus, the

mental representations of political discourse are stored in both levels of memory. According to Chilton (2004), frames are theoretical constructs with some cognitive and neural reality, which are related to the conceptualization of situation types and their expression in language. In the process of conceptualization metaphors play the most important role, so we first need to stop at this point.

Meaning and metaphors

We can distinguish three general approaches to the study of meaning (Taylor, 2003: 186–187):

The language – world approach (semasiology and onomasiology). Meaning is studied as the relationship between linguistic expressions and states of affairs in the world.

The language – internal approach (pragmatic and syntagmatic relations). Meaning is studied in terms of relations between expressions within a language.

The conceptualist approach. The meaning of an expression is equated with a conceptualization in the mind of a language user.

Our analysis here will be based on the conceptualist approach.

Lakoff and Johnson (1980, 1987, 1999) aim to put metaphors at the very centre of cognitive science. Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature (Lakoff, 1980: 454). A metaphor involves a mapping relation between the source and the target domain where reasoning is thought of as if it were a journey in which the traveller moves from one point to the other. The source domain is concrete (daughters being sold, young unprotected girls) and can be perceived directly, while the target domain is more abstract (abuse, morality, women's rights, nationality, etc.).

In this journey, the traveller is supposed to follow the rational line and, if he gets stuck at any point of the rational argument, he will need a companion to help him reach the destination – in our case the leader would appear to be the best and the only companion available (see Lakoff and Johnson, 1999: 236–7). Political discourse, which is a political act, is conceptualized precisely like a metaphorical move or journey. Returning to the text:

This is how we will strengthen the Party, this is how we follow the righteous path, so that the Party, socialism and our people, never have to suffer any head, heart or physical pain [...] We have what it takes, since our Party is

strong, trained and with great revolutionary experience, since the members of our Party are courageous, heroic (E. Hoxha, 1967).

Metaphorical thought is ordinary, and mostly unconscious and automatic (Lakoff, 2008: 82). A further important point about the cognitive theory of metaphor is that metaphorical mappings, which are usually unconscious, are used for reasoning – reasoning about target domains that are ill-understood, vague or controversial (Chilton, 2004: 52). Many cognitive metaphors that Hitler once used in his book *Mein Kampf*, we also find in Enver Hoxha's speeches, such as the one we are analyzing. By referring to the 'enemies of the people' as 'bacteria' or the 'foreigner reactionary', for instance, he emphasizes the source domain and at the same time guides his listeners to the 'correct' inferences, as a good father would do for his children, just in case the reasoning itself is not clear enough.

Roles and deixis

Every discourse has its reality where different role-players act. These role-players or thematic roles (Dowty, 1991; Piker, 1994) are defined by the relations between the parties involved in the discourse and are described using linguistic expressions. The typical linguistic scheme in which these parties are projected in Albanian is Subject-Verb-Object. Consider the example:

'The Party and the nation must take immediate action. To burn the old customs at the stake and behead anyone who breaks the sacred law of the Party in the protection of women's rights. These are very important moral-political problems' (E. Hoxha, 1967).

In political discourse the pronoun has a deictic role, i.e. the pronoun 'we' is used to give identity to the group, party, or coalition. In the first sentence of the text above, this deixis is not sufficient to express the agent, the one that is acting. The Party and the place constitute the subject of the sentence, which assumes a comprehensive character, containing two deictic elements: the Party = us, place = agents outside the party as well as the space deixis. The image schema that is created in the listener's mind is the equation of the Party with the entire population of the country, the comprehensive party, which "should stand on its own feet".

Table 1.

Argument 1 Typical- ly P- Agent, grammat- ical sub- ject, typically a noun phrase Party and nation	Predicate Relation, action, existence, etc., intransi- tive, transitive, or ditran- sitive verb must take immediate action	Argument 2 Typical- ly P- Patient, grammat- ical ob- ject. Typically, a noun phrase	Other arguments e.g., noun, phrase, or preposi- tional phrase	Ad- junct/con- junct e.g., adverbs, participle phrases, conjunc- tions like 'and', 'if'
	to burn at the stake	old cus- toms		and
	behead	anyone		who
	breaks	the sacred law of the Party	in the protection of wom- en's rights	

Here we have two deictic expressions linked to each other in a complex way: the social and the spatial deixis. While social deixis, according to Chilton (2003), arises from social structures and power ties, spatial deixis in political discourse deals with political or geographical spaces or with social groups. The text above is a metaphorical text par excellence where people and other entities are positioned in a clear relationship to each other in space, time and modality.

The text above is a pure linguistic act. It is stripped of the characteristics of political discourse such as: non-explicit expressions, conjectures, linguistic redundancies.

The Strategic Use of Language

From the epistemological aspect it is clear that the speaker is a person who is better acquainted with reality than others. He is the one who reveals the dangers and provides recipes for salvation from them. The speaker takes the side of moral authority and has the legitimacy to do so. Any person who would oppose him is immoral, illegitimate, and should, therefore, be fought. From a deontic point of view, the speaker has the right to exercise such power over the listener because he sides with the 'supreme rationale'.

Table 2.

Text of the head of state	Emotional effects and image schemas
How can a Party cell of the Port in Durrës be considered as vanguard and how can a worker of this port be considered as having a revolutionary position , a person who works well during the day, is distinguished, even praised, yet in the evening, while at home, makes church icons and sells them in the morning to secret believers ?	Creating the image schema in which implications of being held accountable precedes punishment that will follow. Fear that a believer could be discovered.
Party employees are now able to consistently make this necessary diagnosis at work and must come up with the means to heal , on a general scale, for the whole Party, or for individual communists. The cure for specific communists, who do not properly understand, do not apply or who violate the norms, is the study of Marxist-Leninist theory, the revolutionary war and this is part of the Party's general education in this aspect.	Party workers = factory workers, must produce. The party as the only savior for the people. Communists are the doctors who have to cure the disease. The identification of the opposite of the disease that should be fought. Hitler identified Jews with bacteria. The party has the therapy ready.
But during this debate, the value of people will be made clear, how much one person's skin is worth . This is called a revolutionary war in an organization.	The mental schema of the animal whose skin is sold in the market produces strong anxiety and fear. Revolutionary war is "ruthless."
Listen to us, because we will also help you do your work well. But be mindful, if you make a mistake, we will cut your ears off , and if you continue with mistakes, we will chain your legs and there will be no one to protect you . The Party is ours, the government is ours, we are in power, the dictatorship of the proletariat is in power, therefore we will break your head should you involve in evil-doing;	The direct threat escalates to the maximum in which case the individual finds himself completely hopeless when confronted with the power of the Party, namely, the speaker. Strong Fear.
The Albanian Labor Party was established, and it plunged into a terrible war against fascism, which was never seen before by mankind. The fates of our people were at stake and only a Marxist-Leninist party, such as our Party, could save them, as it did save them. So, our	The legitimacy of the government's power finds its roots in the past. The evocation of World War II implies reminiscing about the danger from outside, when the Party emerged 'victorious'. In the same way, it will also win in the future when faced with enemies that are much small-

Party was the sharp, bright, unbreakable and uncompromising sword in the hands of the working class and the Albanian people.	er, such as in the case of an inattentive communist.
Many things happened in the Soviet Union and in countries with popular democracies, which led to socialist regimes being overthrown and the degeneration of their parties. Why didn't this happen to us, too? It is due to the reasons stated above and, again, precisely because of those reasons, nothing is going to happen in the future. Isn't there a continuous imperialist-revisionist coalition against the Labor Party of Albania and socialist Albania?	The parallels drawn with other countries that failed serve to illustrate what could happen in Albania if Party norms are not applied. The Party as the only guarantor of the future. The creation of 'actual' (real, existing) danger from the outside to install uncertainty.

The emotions created, the fear of external and internal danger, are the product of mental representations that stem from discourse, from the selection of words, the structure of sentences. Fear is linked to the fear of invasion (including the historical memory of the Second World War, which, the speaker presumes, some of the listeners still may have) and the fear of domination. There are strong reasons to believe that these emotions are those that evolved in the human mind in order to survive – but any functional component could be used in specific circumstances for specific purposes (Chilton, 2003: 117).

The other as internal enemy: the false communists

To continue further with the analysis, we will use imagology as an auxiliary methodology, specifically by analyzing the way Hoxha legitimizes violence. The other, as a construct, is imagined as an enemy – the embodiment of evil – who is presented with the aid of stereotypes at different levels: regional, cultural, class and ideological (Blažević, 2009; Leerssen, 2007). From the beginning of the text Hoxha makes it clear that the ultimate goal is “the development and revolutionization of the party” according to Marxist ideology, highlighting the problems and “barriers” that “hinder development” and “progress”. He categorizes these barriers as “evil” and “primitive”, which should be completely annihilated and replaced with the “new” i.e. Marxist values. The “barrier” is the Other, which the communists must be willing to destroy. And the enemies

are found everywhere, within the Party and outside the Party, so he provides concrete examples defining who the enemies are and how to overcome the obstacles to progress. The category of this type of prejudice according to Daniel-Henri Pageaux is xenophobic (Pageaux, 2009: 143–144).

Revolutionization of the Party means revolutionization of the communists. These must be steely soldiers, politically clear, courageous, loving, open, honest and, when necessary, strict. They must destroy everything bad, support, organize and fight at the forefront for the progressive initiatives (Hoxha, 1967).

Here Hoxha calls for mobilization, to unmask and mercilessly condemn the anti-communist and anti-socialist enemies within the party. Thus he equates the Party and socialism with the people. Through the categorical comparison of “we” and “they” he successfully outlines the specifics of the Other as internal opponents, whom he calls “non-revolutionaries” and practically “false communists”.

The Other non-communist is presented as a “backward man” who lives a primitive and unhygienic life, without bothering to live a progressive and clean life. With a concrete example, he says that this species lives “on the dung that has filled the stairs of his apartment”. Furthermore, even if he [the false communist] is hygienic, he is not interested in “educating” his neighbours to live a hygienic life, implying that this kind of man is selfish and in contrast to a “true communist”, even if he claims to be a communist.

The false communist is lazy and uninterested, that “(...) who attends the meetings of the neighbourhood, the Front or the professional unions only when called upon.” The other type is the hypocrite, stereotypically categorized as a believer and a bourgeois, who, although a prominent worker by day, “crafts at night icons which he sells for private gain”. By this he implies that the communist is, unlike the non-communists, “clean”, an “educator” of the people, “progressive” and a man who is dedicated to the Party in every respect.

Regional and cultural categorization

The “Other” as the enemy i.e., the “false communists”, is typified by the province, calling attention to the differences between the city and the village. The village and the peasants are described as culturally more “backwards” as a contrast to the communist worker who is progressive. Hoxha com-

compares the culture and lifestyle of the villagers to "garbage" which, according to him, the villagers fanatically preserve. This is visibly shown by concrete examples cherry-picked mainly from the villages and mountainous areas and generalized to apply broadly to the inhabitants of these regions, which suggests that these people continue to engage in backward traditions, such as marrying young girls, contrary to the law. It is no coincidence that he misinterprets the tradition of the bride's dowry – which he calls "the vile habit of selling girls for marriage" – a feature that stereotypes Lezha, a northern and predominantly Catholic city. In doing so, he implies a distinction between the primitive, rural and anti-communist North, and the South, which according to his stereotypes, represents the opposite.

Ideological categorization

The Other as an ideological enemy comes from inside and outside the country. But in fact the enemy is anyone who holds a slightly different view than that of the Party; the idea "with us, or against us" prevails. Hoxha classifies the Other with the "anti-Marxist opposition", which includes liberals, social democrats, Titoists, Trotskyists, and revisionists. He equates them with the invaders that the Party has fought against, i.e. with the Italian fascists, the Germans, the Ballists (National Front), as well as with the "feudal-bourgeois", by which he means the *bey*s and *bayraktars* of the provinces.

There is a constant emphasis on internal enemies who may be "Titoists" as "agents", "traitors", "imperialist-revisionists" and "united imperialists" in reference to the paranoid idea that they are all conspiring against the Party. The Party has enemies of the "khrushchevian traitors", their agencies inside it and outside it, because according to him, fighting them has been a tradition since the war "against the savage fascist coalition" of modern revisionism. Therefore, he categorizes the Other as "modern revisionists and reactionaries".

Legitimizing violence and totalitarianism

Hannah Arendt explained that the disappearance of the private sphere is essential for the creation and existence of totalitarianism. In this way, not only are opponents isolated but also their ideas; in short, the ability to think in such conditions is limited (Arendt, 1973: 367, 474-475). This exact strategy can be traced in the dictator Hoxha's political ideology. The level of legitimization of control of the Party also

has to do with the designation of the public sphere. By "arguing" that these problems, according to him, are "extremely important", the goals of the Party can be achieved merely by merging the sphere of private and public life. Thus, proclaimed private problems and public matters now come within the purview of the Party. Whoever does not act according to the principles of the Party is punished and threatened with death:

"No one can say that this is a personal, private matter when it is related to society, socialist norms, and the Party. (...) the Party is ours; the power is ours, we are in power, the dictatorship of the proletariat is in power, therefore we will break your head in case you do bad things..." (Hoxha, 1967).

In order to increase his legitimacy, Hoxha declared communist laws 'sacred' despite the fact that they stemmed from an atheist ideology. Hoxha considered old customs as primitive and went as far as calling for these traditions to be 'burnt at stakes' and for the beheading of those who violate communist laws.

The Marxist-Leninist ideology itself denies the supreme-transcendent power, which we call God, and by this it reduces the human being to the materialistic category. Nonetheless, when proponents of this ideology seized power, they immediately aimed to imitate God. They even demanded to be honoured as such. Such a demand turns the Party into a pseudo-religious institution, thus exemplifying Cardinal Henry Edward Manning's Hegelian concept that human conflicts are always theological, because humans by their nature are religious (Belloc, 1958: 48; Hegel, 1971: 167). This is why communism by its nature is considered a sort of religion and the Party a sort of Church for militant atheism (see Ryklin, 2008). It is also for this reason that the Party's charter became a sort of 'infallible catechism' — an instruction that ought to be followed dogmatically by everyone.

On the other hand, Hoxha justifies violence by ordering punishments to be meted out only according to the "revolutionary spirit". This is ostensibly in contrast to the spirit of the Inquisition carried out by the Catholic Church, which according to him, condemned individuals in a "meagre" way "for small issues" and politicized revenge and slander on a personal level. Thus, Hoxha claims moral superiority even in regard to punishing his enemies. He adds that: "All these types of [cruel] account requests are

foreign and punishable by our Party, as they are all inspired from the micro-bourgeois point of view”.

Hoxha legitimizes violence by saying that the implementation of these decisions, which are vital for the Party, cannot be carried out in the Party “in a liberal and petty-bourgeois spirit”. He describes the lifestyle and values of the “micro-bourgeois” as “slavery” and continues by depicting the Party as “liberators”. Thus, Hoxha gives the Party an educational role, and violence is legitimized in the name of progress, equality, and freedom (Hoxha, 1967). “Among the innumerable effects caused by the class struggle, the abolition of religion surpasses all limits of imagination [...] The content of this fight does not only demonstrate more clearly than anything else the aggressive physiognomy of the communist state but it also helps to distinguish it from other forms of providing it with an autochthonous character” (Shahu, 2012: 12).

Making the private public

With the disappearance of the private sphere, the next focus is censorship, restriction of thought and the disappearance of the opposition. This is also disguised as unity, freedom and Marxism. The “ideological danger” from potential internal enemies is used as an excuse to suppress critical, ideological thinking, and therefore, to prohibit the existence of opposition even within the Party itself, as this “creates division”. This is expressed in Hoxha’s assertion: “Our party is a monolithic party...[it] does not allow in its ranks factions and oppositions [that are] anti-Marxist, revisionist, Trotskyist, liberal, social democratic, etc.” For him, Marxism-Leninism is an absolute dogma, which should never be opposed and will never tolerate factions that contradict “the unity of Marxist-Leninist thought and action”. He regards opponents within the party as “non-proletarian leftovers” and declares war – by which he means a “great revolutionary school” – to ensure the Party remains as “pure”.

Conclusion

Based on the analysis of Enver Hoxha's speech of February 1967 in accordance with critical discourse analysis, cognitive theory, and imagology as an auxiliary methodology, we have shown how mental representations are structured in order to strengthen the legitimacy of a political leader's power. Conceptual metaphors are based on models that evoke experience that create fear and anxiety.

Accountability is the central theme of the discourse that leads us directly to Goffman's concept of the face. The requirement to give up the “face” legitimizes the dehumanization of the individual, as the first step toward bolstering power. The text is built on the convergence of the three main units of pragmatics/coercion, legitimization/delegitimization, and presentation/misrepresentation to serve the leader's purpose.

The speech has a strong imperative character. All image schemas and emotions serve the speaker's purpose of installing in the listener's mind a revolutionary 'new' world in which he is obliged to believe. They confirm the supreme authority of the speaker, his supreme logic, so that he may give directives for the people that he has legitimized himself. From the grammatical standpoint, the text is mainly built on the third person of the subjunctive in Albanian (with imperative meaning), while some of the verbs are in imperative form. This structure gives a strong coercive character to the whole text, which is linked to legitimacy in the sense that one parameter implies the other.

This type of text leads us to the concept of legitimacy in medieval societies, according to Alfonso and Escalona (2004) in the book, *Building Legitimacy: Political Discourses and Forms of Legitimacy in Medieval Societies*: Far from being a static attribute of power, legitimacy has to be dynamically maintained, competed for, and denied to rivals, and this is valid not only in cases of open conflict, or when rulership is at stake, but also as a continuous process of competition for social power (Alfonso, Escalona, 2004: 12).

Adapting the Chinese Cultural Revolution to the Albanian context (1967) affected society deeply (Nikolla, 2018: 162). A further step of the analysis would be to study the longevity of these mental models in the current political discourse in Albania.

Conflict of Interest and Authorship Confirmation

All authors have participated in conception and design, or analysis and interpretation of the data; drafting the article or revising it critically for important intellectual content; and approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

References

- Alfonso I, Kennedy H, dhe Escalona J (2004) *Building Legitimacy: Political Discourses and Forms of Legitimation in Medieval Societies*.
- Baldi B, Savoia L (2009) *Lingua e comunicazione: La lingua e i parlanti*, Pacini Editore
- Baldi B, Savoia L (2010) Metafora e ideologia nel linguaggio politico, te *Lingua e società*, M. Arcangeli. *Lingua Italiana d'Oggi*, 119–165, Roma: Bulzoni.
- Belloc H (1958) *The Cruise of the "Nona"*. Harmondsworth: Penguin.
- Bregasi M (2015) Përtej debatit politik. Kolegji Dardania, Prishtinë.
- Bregasi M (2020) Saving and Atrocities: Sequence Expansions and Indirectness in Television Interviews, *Human Studies* 43: 89–106
- Chilton P A (1996) *Security Metaphors: Cold War Discourse from Containment to Common House*. New York, Lang.
- Chilton P A, Schäffner C (1997) Discourse and politics. In *Discourse as Social Interaction*, ed.T. van Dijk, 206–230. London: Sage.
- Chilton P A, Schäffner C (2002) *Politics as Text and Talk: Analytic Approaches to Political Discourse*. Amsterdam: John Benjamins.
- Chilton P A (2004) *Analysing Political Discourse: Theory and Practice*. London and New York: Routledge.
- Chomsky N (1968) *Language and Mind*. New York: Harcourt, Brace & World.
- Dijk, T A. van (1987) *Communicating Racism. Ethnic Prejudice in Thought and Talk*. London, Sage.
- Dijk, T A. van (2008) *Discourse and Context: a Sociocognitive Approach*. Cambridge: Cambridge University Press.
- Dijk T A van (2002) Ideology: political discourse and cognition. In *Politics as Talk and Text* eds P. A. Chilton and C. Schäffner. John Benjamins Publishing Company
- David D (1991) Thematic Proto-Roles and Argument Selection, *Language* 67(3): 547–619.
- Fairclough N (1989) *Language and Power*. London and New York: Longman.
- Fairclough N (1992) *Critical Language Awareness*. London: Longman.
- Fairclough N (1995a) *Critical Discourse Analysis: The Critical Study of Language*. London: Longman.
- Fauconnier G (1994) *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge: Cambridge University Press.
- Fauconnier G, Turner M (2002) *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.
- Fetzer A (2002) "Put bluntly, you have something of a credibility problem": Sincerity and credibility in political interviews. In *Politics as Text and Talk* eds P. Chilton and C. Schäffner. John Benjamins Publishing Company.
- Goffman E (1967) *Interaction Ritual: Essays in Face-to-Face Behavior*. New York: Pantheon Books.
- Goffman, E. 1974/1986. *Frame Analysis: An Essay on the Organization of Experience*. Boston: Northeastern University Press.
- Grady J E (1999) A Typology of Motivation for Conceptual Metaphor. Correlation vs. Resemblance. In *Metaphor in Cognitive Linguistics*, eds Nē R. Gibbs Jr. and G. Steen, 79–100. Amsterdam: John Benjamins.
- Grice H P (1975) Logic and conversation. In *Syntax and Semantics: 3: Speech Acts*, eds P. Cole and J. Morgan. New York: Academic Press.
- Habermas J (1984) *The Theory of Communicative Action*. Volume 1. Boston: Beacon Press.
- Hegel G W F (1971) *Early Theological Writings*, trans. T. M. Knox, with an introduction and fragments translated by Richard Kroner, 167 pp. Philadelphia: University of Pennsylvania Press.
- Lakoff G (1993) The Contemporary Theory of Metaphor. In *Metaphor and Thought*, ed. A. Ortony, 202–251. Cambridge: Cambridge University Press.
- Lakoff G (2002) *Moral Politics: How Liberals and Conservatives Think*. Chicago: The University of Chicago Press.
- Lakoff G (2009) *The Political Mind: A Cognitive Scientist's Guide to Your Brain and Its Politics*. New York: Penguin Books.
- Lakoff G, Johnson M (1987) *Metaphors We Live By*. Chicago: Chicago University Press.
- Leerssen J (2007) Imagology: History and method. *Imagology*,. 17–32. Leiden: Brill.
- Nikolla P A (2017) Vështrim antropologjik mbi dhunën ndaj Kishës Katolike në vitin 1967, Shejzat – Pleiades, Rome.

Negau M (2013) *Decoding Political Discourse: Conceptual Metaphors and Argumentation*. Palgrave Pivot.

Pageaux D H (2009) Od kulturnog imaginarija do imaginarnog. [From Cultural Imagery to the Imaginary] in *Kako vidimo strane zemlje. Uvod u imagologiju*, [How we perceive foreign countries. Introduction to imagology] ed. Dukić, D., 143–144. Zagreb: Srednja Europa.

Shahu A (2012) The Communist Precedent of the Elimination of the Religious Elite in Albania (1945–1949).

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

Ratzinger J (2004) *Truth and Tolerance: Christian Belief and World Religions*. San Francisco: Ignatius Press.

Ryklin M K (2008) *Kommunismus als Religion: die Intellektuellen und die Oktoberrevolution*. Verlag d. Weltreligionen.

THE CRIMINAL RESPONSIBILITY OF HIGH-FUNCTIONING AUTISTIC OFFENDERS IN CROATIA

Mladen Bošnjak, Marko Jurjako**, Luca Malatesti***

*Department of Philosophy
Faculty of Humanities and Social Sciences
University of Rijeka
Sveučilišna avenija 4, 51 000 Rijeka, Croatia
email: mladen.bosnjak1@uniri.hr

**Department of Philosophy and Division of Cognitive Sciences
Faculty of Humanities and Social Sciences
University of Rijeka
Sveučilišna avenija 4, 51 000 Rijeka, Croatia
e-mails: mjurjako@ffri.uniri.hr
lmalatesti@ffri.uniri.hr

Abstract:

This paper investigates, from a philosophical perspective, whether high functioning autists are legally responsible for the crimes they may commit. We do this from the perspective of the Croatian legal system. According to Croatian Criminal Law, but also criminal laws adopted in many other countries, the legal responsibility of the person is undermined due to insanity when two conditions are satisfied. The first may be called the incapacity requirement. It states that a person, when committing the crime, suffers cognitive or volitional incapacities or limited capacities that are relevant for exculpation. The second, we may call the mental disorder requirement, which states that these exculpatory incapacities are due to the presence of a mental disorder. In this paper we focus on the incapacity requirement. The review of the studies on executive functions deficits associated with high-functioning autism suggests that many autists should not be held categorically responsible for the crimes they may commit. Nonetheless, we argue that these impairments in executive functions generally do not provide an unqualified excuse.

Key words: High functioning autism; Cognitive and volitional capacities; Criminal responsibility; Executive functions; Duress.

1. Introduction

In this paper we investigate what the appropriate social response to serious offences committed by autistic individuals should be. Autism is thought to be a neurodevelopmental disorder that is characterized by social withdrawal, lack of empathy, and

*This paper is an outcome of the project Responding to Antisocial personalities in a Democratic Society (RAD) that is financed by the Croatian Science Foundation (grant number IP-2018-01-3518). Part of the paper was written at the Institute for Philosophy at the University of Graz. Marko Jurjako would like to thank his hosts in Graz, Norbert Paulo and Lucas Meyer, and especially the Austrian Academy of Sciences for the generous scholarship that funded his research stay.

unusual patterns of behavior, including repetitive behavior and obsessions with peculiar things and rituals (American Psychiatric Association, 2013). Although autism is not significantly correlated with offending and aggressive behavior (see, e.g. Bjørkly, 2009; Heeramun et al., 2017; Yu et al., 2021), we might still wonder whether autists who commit a crime should be held culpable and to what degree (Katz & Zemishlany, 2006).

The question of the responsibility of autistic offenders becomes especially pressing in high profile cases. Let us consider, for instance, the shooting at the Sandy Hook Elementary School in Newtown in 2012 (Connecticut, USA), where Adam Lanza, a twenty-year-old diagnosed with Asperger's syndrome, gunned down and killed 26 people. It seems that one factor which potentially led to this mass shooting was Lanza's unusual preoccupation with guns and violence (see Office of the Child Advocate State of Connecticut, 2014). Some features of autism, such as repetitive behavior and limited interests, seem to present a risk factor for criminal offences (Haskins & Silva, 2006). For instance, these patterns of behavior seem to be associated with thieving and collecting behavior (see Chen et al., 2003), as well as with various forms of stalking (see Stokes & Newton, 2004). Furthermore, the most widespread crime connected with autism is arson (Haskins & Silva, 2006). In one case, an autist burned down a radio station because he could not listen to his favorite channel due to problems with the signal (see Barry-Walsh & Mullen, 2004).

Although the more general and important question of whether offenders who are diagnosed with some form of autism should be held responsible is raised by these dramatic cases, it has received a limited amount of attention in the literature. In the international context, some have argued that, due to the cognitive deficits associated with autism, some autistic individuals will not possess the mental states that are necessary for criminal responsibility (see, e.g. Freckelton, 2011; Katz & Zemishlany, 2006; Volkmar et al., 2021). However, the question of whether autists should be held criminally responsible has not been sufficiently discussed from the perspective of the insanity defense. The insanity defense represents the paradigmatic legal framework that defines psychological incapacities that a person must have to be exculpated for violating the law.

As a partial remedy for such a neglect, we will discuss whether autists should be held criminally responsible from the perspective of Croatian criminal

law. Given that Croatian criminal law conceives criminal responsibility similarly to other countries of the Western hemisphere (Šendula-Jengiđ et al., 2022), the conclusions of our discussion are also relevant for the international context.

The main thesis of this paper is that the criminal accountability of autistic¹ offenders can, in a range of cases, be considered as diminished due to certain cognitive impairments. This means that in criminal proceedings legal defenses that take these mitigating factors into consideration are justified. However, we argue that being diagnosed with autism should not grant full exculpation.

In the rest of the paper, we proceed as follows. In section 2, we offer an overview of autism and explain why we focus on the responsibility of high functioning autists. In section 3, we offer a clarification of the notion of responsibility as codified in Croatian criminal law. Since the law neither describes the exact psychological capacities that are relevant for criminal responsibility nor how their presence or level should be assessed, we argue that there is a need for a *bridging principle* that could connect our folk understanding of responsibility and scientific evidence that might have a bearing on this issue. In section 4, we argue that the executive function theory of responsibility offered by William Hirstein, Katrina Sifferd, and Tyler Fagan (2018) supplies a satisfactory bridging principle between the legal understanding of responsibility capacities and the relevant scientific evidence that might be used to test the presence of these capacities. In section 5, we argue that these deficits in executive functions do not completely exculpate high-functioning autists because they have sufficient psychological capacities to overcome cognitive and volitional peculiarities that would otherwise reduce their culpability.

2. Autism spectrum disorder

Autism is a neurodevelopmental condition that is characterized by stereotypical and repetitive behavior, a limited range of interests, and deficits in social interaction and communication (American Psychiatric Association, 2013). The severity of these deficits varies among autists. Thus, in the prominent psychiatric manuals, autism tends to be defined as a

¹ Some studies indicate that the person-first language increases stigma (Gernsbacher, 2017). To avoid stigma, we adopt the identity-first language instead of person-first language. Thus, we will write "autist" or "autists" instead of "a person with autism" or "individuals with autism."

disorder whose symptoms and severity fall on a spectrum.² In the latest edition of the *Diagnostic and Statistical Manual* (DSM-5, American Psychiatric Association, 2013), autism spectrum disorder is defined by symptoms that are divided into two classes: social communication and restricted and repetitive behaviors. Symptoms that characterize the social communication aspect of autism include deficits in language comprehension, delay in learning to speak, avoidance of eye contact, and lack of understanding of social cues. Restricted and repetitive behavior is characterized by a tendency to focus on one topic, repeating words or phrases, and repetitive movements.

The DSM-5 includes 3 levels of severity in autistic symptoms that show the level of disability that autists face in performing daily activities (APA, 2013, 299.00; F84.0). Level 1 refers to those individuals who need some support to perform daily routines. For instance, they might have problems in social communication and be socially withdrawn but have no serious problems in living independently and supporting themselves. In earlier versions of the DSM, Asperger syndrome was a separate category that would fall under level 1 severity of symptoms. In the DSM-5, Asperger syndrome was removed, and all conditions related to autism were subsumed as a constellation of traits under autism spectrum disorder. People who have level 1 traits on the autism spectrum could be said to have high-functioning autism.

Level 2 refers to individuals who need substantial support. Those would be individuals who have more severe problems in social communication and, due to repetitive behavior and narrowed interests, have difficulties in adapting to changing social situations. To overcome these challenges, they need substantial support from other neurotypical individuals.

Level 3 refers to individuals who need very substantial support to the point of not being able to live on their own or to autonomously perform most eve-

ryday activities. Those would be individuals who have severe difficulties in communication with other people, do not speak, and have lower than average IQ. Given these difficulties in social communication and impairments in everyday activities, level 2 and 3 autists would fall on the side of the autism spectrum that might be called low-functioning autism.

The criminal accountability status of low-functioning autists seems to be clear-cut. Due to severe behavioral, psychological, or social impairments, low-functioning autists are unable to live on their own, and thus tend to be taken care of by family members or specialized institutions. Moreover, due to intellectual disabilities, inflexibility of behavior, and extremely diminished capacities for social interaction it makes sense to think that low-functioning autists by default should be held less accountable for potential violations of the law (Feather, 2016).

Thus, in this paper, we will mostly discuss the legal status of high-functioning autists. It is standardly supposed that high-functioning autists have capacities that enable them to be well-integrated into the fabric of society. This means that despite certain behavioral and cognitive peculiarities they can engage with other individuals and form relatively normal connections with other people (see, e.g. Lekić Barunčić, 2021). Nonetheless, the question of the responsibility of high-functioning autists is perplexing. This group includes individuals with mild autistic traits and without intellectual impairments. Thus, at least at first glance, they appear to have capacities needed for moral and legal responsibility. However, at the same time they manifest certain behavioral difficulties and problems with social interactions that might suggest that they should not be held entirely responsible for criminal acts they may commit.

In the next section, we will introduce the notion of criminal responsibility as construed in Croatian criminal law. This will provide us with the framework for discussing the responsibility status of high-functioning autists.

3. Criminal Responsibility in Croatian Criminal Law

According to the Croatian criminal code, there are four components of culpability: (1) accountability, (2) intention or negligence, (3) awareness of illegality or possibility of that awareness and (4) absence of excusing reasons (Grozđanić et al., 2013, p. 131). Here we consider whether autists have the general capacities necessary to be criminally respon-

² It should be noted that whether autism is a disorder depends on how the notion of disorder is understood and then applied in psychiatric manuals (see, e.g. Biturajac & Jurjako, 2022). Thus, depending on which account one favors, the question of whether autism is a disorder could receive a different verdict (see, e.g. Wakefield et al., 2020, see also footnote 3). Moreover, this question is important because to a large extent it determines whether we should try to alter autistic traits via medical means or treat them as normal variations (for discussion of medicalization, see, e.g. Gagné-Julien, 2021; Lekić Barunčić, 2021; Stegenga, 2021).

sible. Thus, we will mostly focus on the first component of culpability.

According to the Croatian criminal code, defendants can be exculpated from criminal responsibility if they are mentally incapable:

A mentally incapable person is a person who at the time of an unlawful conduct is incapable of appreciating the meaning of his or her conduct or of exercising control over his or her will due to mental illness, temporary mental disorder, insufficient mental development or some other severe mental disorder (Croatian Criminal Code 2011, article 24, paragraph 2).

Croatian criminal law recognizes mental incapacity as a mental disorder that causes serious impairments in what H.L.A Hart (1968) has termed “capacity responsibility”. Capacity responsibility includes two abilities. One is the capacity to understand the requirements of the law, which in the Croatian criminal code is expressed as the ability to appreciate the meaning of one’s conduct. This element of capacity responsibility is usually dubbed as the cognitive requirement. The second is the ability to control one’s behavior by following one’s understanding of the requirements of the law. In the Croatian criminal code this is expressed as the ability to control one’s will. Thus, according to the Croatian criminal code, a person is not accountable if they, due to a mental disorder, cannot understand the nature of their actions and/or do not have the ability to control their actions in light of that understanding.

Given that accountability is tied to an individual’s cognitive capacities, and is therefore a matter of degree, the Croatian legal system, as many others, recognizes diminished accountability as a condition in-between accountability and unaccountability (see Croatian criminal code, article 26). It is important to note that diminished accountability is not by itself a basis for mitigation of the sentence, although it can be considered as a mitigating circumstance (Grozđanić et al., 2013, p. 139). Only substantially diminished accountability is a facultative base for mitigation of the sentence, i.e., imposing a sentence above the prescribed minimum for a criminal act in question (Grozđanić et al., 2013, p. 140). This rule does not apply in the case of self-induced unaccountability because in such a case the offender has intentionally or negligently brought him or herself to “a state in which he or she was incapable either of appreciating the meaning of his or her conduct or of exercising

control over his or her will” (Croatian Criminal Code 2011, article 25). Paradigmatic examples of self-induced unaccountability are alcohol and drug abuse (see Grozđanić et al., 2013, p. 142).

High-functioning autists seem to satisfy the mental disorder requirement for exculpation – at least in the sense that high-functioning autism is encompassed by autism spectrum disorder in most diagnostic manuals.³ Nevertheless, even if we accept that high-functioning autism is a mental disorder this is not sufficient to consider high-functioning autists as less responsible. Croatian criminal code requires that the mental disorder should influence the capacity responsibility of the defendant. Thus, if high-functioning autists have diminished responsibility, this claim should be justified by explaining how cognitive or volitional deficits associated with high-functioning autism undermine their capacity responsibility (Grozđanić et al., 2013; see, also Freckelton, 2011; Malatesti et al., 2020). Thus, to answer this question, we should investigate what the cognitive and volitional deficits associated with high-functioning autism are and whether they substantially diminish capacity responsibility.

This question cannot be straightforwardly answered, though. The Croatian criminal code does not specify what the exact mental capacities that underpin criminal responsibility are, and, thus, how their presence or absence can be empirically investigated. This issue tends to be relegated to forensic psychiatrists. Thus, to solve this problem we would need to see how forensic psychiatry, both in its theoretical and practical dimensions, addresses the issue of identifying exculpatory incapacities (see, e.g. St. Stoyanov, 2018; O’Sullivan, 2018). However, rather than investigating the concrete empirical details of how exculpatory incapacities are distinguished in specific cases, we are more interested in discussing conceptual and foundational aspects of the general problems involved in forensic science. These issues become important especially in controversial cases concerning whether an entire class of individuals,

³ However, it should be noted that this view is challenged in various ways by the proponents of the neurodiversity movement (see, e.g. Meyerding, 2014; Armstrong, 2015; Ortega, 2009). Given that the question of whether autism is a mental disorder is not important for the present paper, we will not further discuss it. For a philosophical analysis of this question, see, e.g. Hughes (2021), Lekić Barunčić (2021), Nelson (2021), and Wakefield, Wasserman, and Conrad (2020).

such as people diagnosed with autism, are criminally responsible.

More specifically, there is a general problem of “interfacing” the notions of psychological capacities that underpin criminal responsibility and the empirical sciences of the mind that are used to prove the presence or absence of the relevant psychological capacities (see, e.g. Buckholtz et al., 2016; Francken & Slors, 2018; St. Stoyanov, 2018). This so-called “interface problem” stems from the following considerations. On the one hand, our understanding of criminal responsibility is based on common sense psychological concepts such as beliefs, wishes, and intentions. The criminally accountable person is someone who has normal beliefs about her social and physical surroundings and can act following her desires and plans (Malatesti et al., 2020). Thus, at this level of explanation, we talk about rational individuals who act on their reasons. On the other hand, science relies on causal explanations spanning from cognitive mechanisms underpinning subconscious and impersonal perceptual and reasoning processes to neuroscientific explanations that are framed in terms of brain mechanisms underlying cognitive, emotional, and motor capacities (Janik & Próchnicki, 2018). The problem is how to translate and connect the (neuro)scientific data with the folk-psychological concepts in a way that would make them relevant for determining whether a person is criminally responsible for her actions (St. Stoyanov, 2018).

Of course, the interface problem is not really a problem if the scientific descriptions of human psychological capacities clearly show that these capacities are impaired in a way that they undermine capacity responsibility. This may be so in the case of low functioning autists. Even a raw description of their psychological peculiarities is enough to grant complete or partial exculpation. However, the interface problem becomes more salient when the cognitive deficits are more subtle and not visible on the commonsense psychological level of description, and the (neuro)scientific accounts of psychological capacities are needed to solve this issue (see, e.g. Moore 2016; Janik & Próchnicki, 2018).

In the next section, we explore, thus, whether a general bridging principle could connect satisfactorily the legal conceptions of the capacities underpinning criminal responsibility and those capacities examined from a (neuro)scientific perspective. Hirstein, Sifferd, and Fagan (2018) offer a plausible solution to this difficulty. They redescribe the rele-

vant notion of criminal responsibility through the lens of the neuroscientific notion of executive function.

4. The executive function theory of responsibility

According to Hirstein and colleagues (2018, p. 18), executive functions underpin cognitive and volitional capacities, which are relevant for criminal responsibility. Accordingly, mental disorders that undermine criminal responsibility would be those that in the relevant ways impair executive functions.

On the cognitive level, “executive functions” refer to processes that play a key role in dealing with new situations and complex tasks, unlike everyday routines, which we tend to perform automatically. At this higher level, they are thought to be composed of processes underlying working memory, attention, and inhibition. Thus, they refer to processes that enable us to control our actions. In general, executive functions enable us to make plans, form intentions and habits, and alter them, regulate behavior by considering inputs from emotions, memory, senses, and motivational states (Hirstein et al., 2018, p. 18). It should be noted that the ability to use executive functions may change during one’s life course. Development, education, interaction with peers, and family can improve performance of executive functions. Furthermore, impairments in executive functions may be improved by therapeutic intervention, medication, surgery, or cognitive and behavioral training (Otero & Barker, 2014).

On the implementational level, executive functions are thought to be realized in the frontal lobes of the brain. These include areas such as the dorsolateral prefrontal cortex, anterior cingulate, and orbit medial prefrontal cortex. Neurons from these areas also project to subcortical regions that play important roles in processing emotional and motivationally salient stimuli (Otero & Barker, 2014).

Preserved executive functions are extremely important for normal functioning and behavioral control. For instance, executive dysfunction is related with poor behavioral control, rigidity in thought, failure in planning ahead and starting action, difficulties in focusing attention, and so on. Deficits in such processes are associated with maladaptive and impulsive behavior that underlie externalizing disorders, and they might lead to severe forms of antisocial behavior (see, e.g. Brazil et al., 2018).

According to Hirstein, Sifferd, and Fagan, the connection between executive functions and legal

responsibility can be spelled out as follows. A person, who might be called Jo, is responsible only if:

1. Jo has a minimal working set of executive functions (MWS), and
2. Jo performed the act or omission, and/or caused the consequence, or failed to act to prevent it, and
3. Jo's executive processes either played an appropriate role in bringing about the action, omission, or consequence or should have played an appropriate role in preventing it (Hirstein et al., 2018, p. 87).

According to the first condition, the person needs to have a certain degree of overall executive functioning to be the proper target of responsibility assessments. Some people are naturally better at performing tasks that use executive functions than others (e.g., in planning ahead, having greater working memory capacity, and so on). Nonetheless, the idea is that criminal responsibility presupposes that people minimally have a basic level of executive functions that are necessary for adaptive functioning in a society and performing everyday activities. Hirstein and colleagues (2018, p. 57) use an analogy with cars: despite differing in many respects all cars, regardless of their differences, should satisfy minimum safety standards to be driven.

According to condition 2, there should be a suitable relation between an offender's acts or omissions and their mental states (Hirstein et al., 2018, p. 87). Here the idea is that we are responsible both for bad outcomes of our actions and for those bad outcomes that our actions could have prevented (Hirstein et al., 2018, p. 64). For example, if person A knows that person B plans to murder person C, person A can inform the police about B's plans and thus prevent the murder. If she does not do that, she can be held liable for not alerting the relevant institutions about the predicted crime.

Condition 3 refers to relations that need to obtain between an offender's executive processes and actions in order to be criminally responsible for them. It says that in "positive" cases, performance of one or more executive functions must causally lead to an action (Hirstein et al., 2018, p. 66). Causal history is especially important; the level of culpability is higher if someone has been driving at full speed and intentionally ran into a crowd than if the same effect is caused by accidentally pressing the wrong pedal due to lack of focus. The person in the first case is more blameworthy because they planned

the action, with the intention of hurting or killing somebody, and finally executed that intention.

In this paper we are addressing the question of the responsibility of high-functioning autists in general. Thus, we will focus on the question of whether autists satisfy the first criterion of responsibility. The remaining two criteria concern specific circumstances in singular cases and therefore are not relevant for the purposes of the paper. In the next section, we will examine the scientific data that might be relevant for deciding whether high-functioning autists suffer from impairments in executive functions that might reduce their culpability.

5. Should high-functioning autists be legally excused?

Studies show that high-functioning autism is associated with impairments in executive functions that underpin cognitive and volitional capacities as captured by Croatian criminal law. Autists manifest difficulties with working memory, as demonstrated by the fact that they may have trouble following instructions, despite correctly understanding them (Geurts et al., 2014, p. 127). This might explain problems in social interactions because working memory plays a key role in storing, processing, and noticing different social cues, such as facial expressions, tone of voice and body language (Geurts et al., 2014, p. 127; Causton-Theoharis et al., 2009). Autism also seems to be associated with cognitive inflexibility, which is shown in stereotypical and repetitive behavior (Monsell, 2003).

Autism is also associated with reality monitoring deficits (see, e.g., Grainger & Williams 2014; Cooper et al., 2016), that affect the capacity to differentiate "between internal and external sources of information" (Cooper et al., 2016, p. 2186). Thoughts and imagination exemplify the internal sources of information, while perceptual processes are the external sources of information. Some studies show that autists have problems in distinguishing between these two sources of information (Simons et al., 2008; Cooper et al., 2016).

These deficits in executive functions may contribute to and explain cases of unusual criminal behavior committed by some autists. For instance, Murrie and colleagues (2002) mention the case of a man who set several houses on fire in revenge for being bullied in childhood. However, he was not aware that the victims who lived in those houses had no relation to the people who had bullied him. He set

fire to those houses simply because to him they appeared like the houses of his abusers.

Some authors, such as, most notably, Simon Baron-Cohen (1995), argue that autists have serious deficits in the capacities underlying theory of mind. Theory of mind denotes the capacity to ascribe mental states such as beliefs, desires, and emotions to others based on their behavior in particular contexts. For example, if we see someone running toward a bus early in the morning, we might explain their behavior by inferring that they do not want to be late for work and that by catching the bus they will satisfy this desire. Similarly, if we see someone exercising in the gym, we will likely think that they want to be healthy and that they believe that by exercising they can achieve this goal. Thus, theory of mind enables us to explain other people's behavior in terms of mental states, such as desires and beliefs, that we ascribe to them. According to the theory of mind deficit hypothesis, autists have problems predicting and explaining other people's behavior based on the mental states we normally ascribe to them.

The most compelling evidence for this hypothesis is based on how autists perform on the false belief task (Baron-Cohen et al., 1985). In the most common version of the task, participants see a character, Sally, who plays with a marble and, once she is finished playing, puts the marble in a basket and leaves the room. After that, another character enters the scene, Anne, who takes the marble from the basket and puts it in a box that is placed near the basket. Participants are asked to predict where Sally will look for the marble once she comes back into the room. Participants who have normal theory of mind capacities answer that Sally will search for the marble in the basket. This answer is expected because they saw Sally leave the marble in the basket and they know that Sally did not see Anne move the marble into the box. Participants who respond in this way can understand that Sally's behavior is motivated by her *false* belief that the marble is in the basket. Studies show that neurotypical individuals from the age of 4 can solve this task (Baron-Cohen et al., 1985). This shows that kids from the age of 4 grasp the concept of belief and can predict and explain people's behaviors by attributing mental states to them and interpreting things from their point of view. However, studies show that autists have problems solving this task (Baron-Cohen et al., 1985; Girli & Tekin, 2010). When they are asked where Sally will look for the marble, they tend to say that she will look in the box where Anne has put it. This

shows that autists have deficits in the theory of mind because of which they have problems in using this knowledge to predict and explain other people's behavior.

Deficits in the theory of mind associated with autism might explain failures to consider the perspective of others, which is extremely relevant for healthy social relations and might have a role in criminal behaviors. For instance, such deficits can explain some of the sex crimes committed by autists. It seems that autists who sexually abuse their victims tend to dehumanize them and treat them as pure objects. This might be a consequence of a failure to take the perspective of other people into consideration and to understand their cognitive and emotional states (see Silva et al., 2004).

Regarding volitional capacities, autists seem to exhibit planning difficulties, such as making a schedule of daily duties or managing a household (Geurts et al., 2014, p. 130; see also Bennetto et al., 1996; Lopez et al., 2005; Griebeling et al., 2010). Planning problems might also contribute to difficulties in social interactions and communication. Autists often need help from their family to plan social activities and to keep social contacts (Geurts et al., 2014). Moreover, some studies suggest that autism is related to decreased error processing and lower correction of behavior after committing a mistake (Sokhadze et al., 2010). Deficits in error-processing and response monitoring might disable autists from accurately assessing a behavioral mistake and learning from it, which in turn may also explain the rigid and repetitive behavior that is distinctive of autism (Sokhadze et al., 2010). Autists also suffer from inhibitory problems. They are often not able to control their behavior, although they are aware of its negative outcomes, as confirmed in studies that show that autists have problems in disregarding irrelevant (interfering) details when solving a task (see Adams & Jarrold, 2012; Christ et al., 2011; Geurts et al., 2008). Such difficulty in inhibiting prepotent action might also explain difficulties in social interactions and repetitive behavior (Geurts et al., 2014, pp. 125-126).

Although there is evidence for associating high-functioning autism with impairments in executive functions, these impairments do not seem to be of such a degree as to call for exculpation (see Richman, 2018). First, many of these studies have not been replicated. For instance, some studies show that reality monitoring is impaired in autism (Lind & Bowler, 2009; Russell & Jarrold, 1999), while others

show that reality monitoring is intact (Hill & Russell, 2002; Zalla et al., 2010). Also, some studies report no deficits in behavior monitoring and learning from errors (see Hill & Russell, 2002). Similarly, there is inconsistency between studies on cognitive inflexibility associated with autism (Geurts et al., 2014, pp. 129-130; see also Geurts et al., 2009). Some studies suggest impairments in cognitive flexibility in autism (see, e.g., Ozonoff et al., 2004; Reed & McCarthy, 2012), while others show that there are no such impairments (see, e.g., Corbett et al., 2009; Happé & Frith, 2006). Inconsistencies in these studies show that either high-functioning autists do not suffer from such problems, or they are not of such a degree as to be reliably attributed to an average high-functioning autist.

Second, despite potential deficits in executive functions, it seems that high-functioning autists on average have a minimally working set of executive functions that underlie capacity responsibility. Most high-functioning autists manage to live independently and be integrated into the social fabric of society (American Psychiatric Association, 2013). Moreover, to fit in with the rest of their society, autists tend to use various strategies for social camouflaging (see, e.g., Boren, 2017; Hull et al., 2017). This shows that they tend to be aware of their cognitive or volitional impairments. Such awareness makes them responsible because they can use this knowledge to compensate for their impairments – especially those that pertain to theory of mind deficits. This might be achieved in several ways. For instance, in social interactions, this might involve faking facial expressions, learning social scripts in advance, keeping eye contact by force, and simulating empathic concern for other people, even if they do not care about them (Boren, 2017; Hull et al., 2017; Lai et al., 2019).

Nonetheless, it might be argued that even if exercising these capacities would be possible on some occasions, most often it would be so straining for autists as to undermine or, at least, diminish their responsibility. For instance, some claim that using such compensatory strategies may negatively affect autists' sense of identity and well-being, causing frustration and loneliness (see Hull et al., 2017). Thus, it might seem that autists, despite having the relevant capacities, should often not be held responsible for their actions when this depends on using abilities and heuristics for socially adaptive behavior that might cause them harm.

Although it might be extremely hard and unpleasant for autists to use such compensatory strategies, still it does not follow that they should be excused. Just because some action is difficult and unpleasant does not mean that by law one should be excused for not performing it. For instance, in a corrupt society it might be difficult not to be a corrupt young politician given the social structures and social expectations set up by senior politicians. Nonetheless, succumbing to the pressure of others is not *ipso facto* an excusing condition. The difficulty to conform one's actions to the law must involve an *incapacity* or *reduced* capacity to control one's actions (Grozđanić et al., 2013; see, also Šendula-Jengiće et al., 2022).

There might be another way to argue that autists face such difficulties in overcoming their default cognitive and behavioral deficits as to be held less responsible for potential criminal offences. It might be suggested that using compensatory strategies would often be experienced by autists as duress. They can overcome many cognitive and behavioral difficulties but for them doing so would be like forcing somebody to commit an atrocious act under pressure of retaliation. In such a situation, the person does commit the atrocious act, but they are not responsible for it because they were compelled to do so under threat.

The plausibility of such reasoning depends on how we understand duress in this context. Since autists are not forced by the peculiarities in their executive functions to violate moral or criminal codes the standard legal notion of duress does not apply in this case. However, we might still be expecting too much of the average autist by requiring that they continuously effortfully adjust their actions to social expectations. Because of the challenge this presents for them they may feel like they are complying with legal and social expectations under duress. However, if this situation is characterized as duress, then this notion should be understood as a *choice* between two options that, although undesirable, we can still be held responsible for.

In this regard, in a closely related context, Carl Elliot (1996) offers an account of duress that might be appropriated for our purposes. According to Elliott (1996, Ch. 3), duress is a situation in which "the person is faced with the choice between (1) acting on desires that he finds morally repellent and shameful and, (2) refraining, which causes him considerable distress" (Elliott, 1996, p. 48). Under this interpretation, we might understand high-functioning

autists as often facing a choice between either acting on desires that lead to criminal offences or investing considerable effort in order to refrain from acting on such desires, even to the point of causing distress to themselves (see also Petrolini, 2020). It is important to note that even in this case duress should be understood as “a choice between undesirables” (Elliott, 1996, p. 46) despite the fact that “the person who has acted under duress typically would strongly prefer not to be faced with the choice [they] must make” (Elliot, 1996, p. 47). Thus, duress understood in this way retains “at least some element of decision” (Elliot 1996, p. 46) and entails at least some degree of accountability.

Thus, if duress is understood in those terms, we might concede that the responsibility of autists will often be diminished. However, it will not be entirely absent. Autists might be seen as forced to choose between two bad alternatives; on the one hand, they can behave in accordance with their autistic nature, thereby risking breaking the law, or, on the other hand, they can try to suppress their autistic behavior and potentially suffer because of it. Whatever they do, they are responsible because they have sufficiently preserved executive functions to make a choice between at least two options that are available to them.

6. Conclusion

In this paper, we have discussed the accountability of high-functioning autists. Based on the executive function theory of responsibility offered by Hirstein, Sifferd, and Fagan (2018), we have argued that despite impairments in executive functions, autists can only be regarded as having diminished responsibility. This is because high-functioning autists are aware of their cognitive impairments and can rely on preserved executive functions to partially compensate for their impairments. We have also argued that because in many situations acting against their behavioral dispositions might be experienced by autists as duress, this might be taken as a mitigating circumstance when considering whether autists should be held completely responsible for their actions. However, even in this case we have argued that experiencing duress should be construed as facing a choice between two options, albeit unpleasant ones: they can try to overcome their cognitive deficits, which might involve efforts to conform to neurotypical reasoning and behavioral styles, or they may refuse to invest such efforts, thereby increasing their chances of violating the law.

References

- American Psychiatric Association. (2013) *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. Arlington, VA: American Psychiatric Association.
- Adams N C, Jarrold C (2012) Inhibition in autism: Children with autism have difficulty inhibiting irrelevant distractors but not prepotent responses. *Journal of Autism and Developmental Disorders* 42(6): 1052–1063. <https://doi.org/10.1007/s10803-011-1345-3>.
- Armstrong T (2015) The myth of the normal brain: Embracing neurodiversity. *AMA Journal of Ethics* 17(4): 348–352. <https://doi.org/10.1001/journalofethics.2015.17.4.msoc1-1504>.
- Baron-Cohen S (1995) *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press.
- Baron-Cohen S, Leslie A, Frith U (1985) Does the autistic child have a ‘theory of mind?’ *Cognition* 21: 37–46.
- Barry-Walsh J B, Mullen P E (2004) Forensic aspects of Asperger’s syndrome. *The Journal of Forensic Psychiatry & Psychology* 15(1): 96–107. <https://doi.org/10.1080/14789940310001638628>
- Bennetto L, Pennington B F, Rogers S J (1996) Intact and impaired memory functions in autism. *Child Development*, 67(4): 1816–1835.
- Bermúdez J L (2005) *Philosophy of Psychology: A Contemporary Introduction*. London: Routledge.
- Biturajac M, Jurjako M (2022) Reconsidering harm in psychiatric manuals within an explicationist framework. *Medicine, Health Care and Philosophy*. <https://doi.org/10.1007/s11019-021-10064-x>
- Bjorkly S (2009) Risk and dynamics of violence in Asperger’s syndrome: A systematic review of the literature. *Aggression and Violent Behavior* 14(5): 306–312. <https://doi.org/10.1016/j.avb.2009.04.003>
- Boren R (2017) Autistic burnout: The cost of masking and passing. Retrieved March 2, 2022, from <https://boren.blog/2017/01/26/autistic-burnout-the-cost-of-coping-and-passing/>.
- Brazil I A, van Dongen J D M, Maes J H R, Mars R B, Baskin-Sommers A R (2018) Classification and treatment of antisocial individuals: From behavior to biocognition. *Neuroscience & Biobehavioral Reviews* 91: 259–277. <https://doi.org/10.1016/j.neubiorev.2016.10.010>

- Buckholtz J W, Reyna V F, Slobogin C (2016) A neuro-legal lingua franca: Bridging law and neuroscience on the Issue of self-control. *Mental Health Law & Policy Journal*.
<https://ssrn.com/abstract=2788178>
- Causton-Theoharis J, Ashby C, Cosier M (2009) Islands of loneliness: Exploring social interaction through the autobiographies of individuals with autism. *Intellectual and Developmental Disabilities* 47(2): 84–96. <https://doi.org/10.1352/1934-9556-47.2.84>
- Chen P S, Chen S J, Yang Y K, Yeh T L, Chen C C, Lo H Y (2003) Asperger's disorder: A case report of repeated stealing and the collecting behaviours of an adolescent patient. *Acta Psychiatrica Scandinavica* 107(1): 73–75; discussion 75–76. <https://doi.org/10.1034/j.1600-0447.2003.01354.x>
- Christ S E, Kester L E, Bodner K E, Miles J H (2011) Evidence for selective inhibitory impairment in individuals with autism spectrum disorder. *Neuropsychology* 25(6): 690–701. <https://doi.org/10.1037/a0024256>
- Cooper R A, Plaisted-Grant K C, Baron-Cohen S, Simons J S (2016) Reality monitoring and metamemory in adults with autism spectrum conditions. *Journal of Autism and Developmental Disorders* 46: 2186–2198. <https://doi.org/10.1007/s10803-016-2749-x>
- Corbett B A, Constantine L J, Hendren R, Rocke D, Ozonoff S (2009) Examining executive functioning in children with autism spectrum disorder, attention deficit hyperactivity disorder and typical development. *Psychiatry Research* 166(2–3), 210–222. <https://doi.org/10.1016/j.psychres.2008.02.005>
- Croatian Criminal Code (2011) The Croatian Parliament. Retrieved March 2, 2022. https://www.legislationline.org/download/id/7896/file/Croatia_Criminal_Code_2011_en.pdf
- Elliott C (1996) *The Rules of Insanity: Moral Responsibility and the Mentally Ill Offender*. SUNY Press.
- Feather K A (2016) Low functioning to high functioning autism: A prescriptive model for counselors working with children across the spectrum. Article 11. *American Counseling Association*. https://www.counseling.org/docs/default-source/vistas/article_11d2bf24f16116603abcacff0000bee5e7.pdf?sfvrsn=4ca9442c_4
- Francken J C, Slors M (2018) Neuroscience and everyday life: Facing the translation problem. *Brain and Cognition* 120: 67–74. <https://doi.org/10.1016/j.bandc.2017.09.004>
- Freckelton I (2011) Asperger's disorder and the criminal law. *Journal of Law and Medicine* 18(4): 677–694.
- Gagné-Julien A-M (2021) Wrongful medicalization and epistemic injustice in psychiatry: The case of premenstrual dysphoric disorder. *European Journal of Analytic Philosophy* 17(2) (S14): 5–36. <https://doi.org/10.31820/ejap.17.3.3>
- Gernsbacher M A (2017) Editorial perspective: The use of person-first language in scholarly writing may accentuate stigma. *Journal of Child Psychology and Psychiatry* 58(7): 859–861. <https://doi.org/10.1111/jcpp.12706>
- Geurts H, de Vries M, Bergh S (2014) Executive Functioning Theory and Autism. In: *Handbook of Executive Functioning*, eds Goldstein S., Naglieri J. New York, NY: Springer. https://doi.org/10.1007/978-1-4614-8106-5_8
- Geurts H M, Begeer S, Stockmann L (2009) Brief report: Inhibitory control of socially relevant stimuli in children with high functioning autism. *Journal of Autism and Developmental Disorders* 39(11): 1603–1607. <https://doi.org/10.1007/s10803-009-0786-4>
- Geurts H M, Luman M, van Meel C S (2008) What's in a game: The effect of social motivation on interference control in boys with ADHD and autism spectrum disorders. *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 49(8), 848–857. <https://doi.org/10.1111/j.1469-7610.2008.01916.x>
- Girli A, Tekin D (2010) Investigating false belief levels of typically developed children and children with autism. *Procedia - Social and Behavioral Sciences* 2: 1944–1950. <https://doi.org/10.1016/j.sbspro.2010.03.261>
- Grainger C, Williams D M, Lind S E (2014) Metacognition, metamemory, and mindreading in high-functioning adults with autism spectrum disorder. *Journal of Abnormal Psychology* 123(3): 650–659. <https://doi.org/10.1037/a0036531>
- Griebing J, Minshew N J, Bodner K, Libove R, Bansal R, Konasale P, Keshavan M S, Hardan A (2010) Dorsolateral prefrontal cortex magnetic resonance imaging measurements and cognitive performance in autism. *Journal of Child Neurology* 25(7): 856–863. <https://doi.org/10.1177/0883073809351313>
- Grozđanić V, Škorić M, Martinović I (2013) *Kazneno pravo: Opći dio [Criminal Law: The General Part]*. Rijeka: Faculty of Law.

- Happé F, Frith U (2006) The weak coherence account: Detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders* 36(1): 5–25. <https://doi.org/10.1007/s10803-005-0039-0>
- Hart H L A (1968) *Punishment and Responsibility: Essays in the Philosophy of Law*. New York: Oxford University Press.
- Haskins B G, Silva J A (2006) Asperger's disorder and criminal behavior: Forensic-psychiatric considerations. *The Journal of the American Academy of Psychiatry and the Law* 34(3): 374–384.
- Heeramun R, Magnusson C, Gumpert C H, Granath S, Lundberg M, Dalman C, Rai D (2017) Autism and convictions for violent crimes: Population-based cohort study in Sweden. *Journal of the American Academy of Child & Adolescent Psychiatry* 56(6): 491–497.e2. <https://doi.org/10.1016/j.jaac.2017.03.011>
- Hill E L, Russell J (2002) Action memory and self-monitoring in children with autism: Self versus other. *Infant and Child Development* 11(2): 159–170. <https://doi.org/10.1002/icd.303>
- Hirstein W, Sifferd K, Fagan T (2018) *Responsible Brains: Neuroscience, Law, and Human Culpability*. Cambridge, Massachusetts: The MIT Press.
- Hughes J A (2021) Does the heterogeneity of autism undermine the neurodiversity paradigm? *Bioethics* 35: 47–60. <https://doi.org/10.1111/bioe.12780>
- Hull L, Petrides K V, Allison C, Smith P, Baron-Cohen S, Lai M-C, Mandy W (2017) “Putting on my best normal”: Social camouflaging in adults with autism spectrum conditions. *Journal of Autism and Developmental Disorders* 47(8): 2519–2534. <https://doi.org/10.1007/s10803-017-3166-5>
- Janik B, Próchnicki M (2018) Naturalizing the subjective side of the crime: A few introductory remarks on the role of consciousness in criminal law based on American and Polish examples. *Balkan Journal of Philosophy* 10(1): 37–44. <https://doi.org/10.5840/bjp20181015>
- Katz N, Zemishlany Z (2006) Criminal responsibility in Asperger's syndrome. *The Israel Journal of Psychiatry and Related Sciences* 43(3): 166–173.
- Lai M-C, Lombardo M V, Chakrabarti B, Ruigrok A N, Bullmore E T, Suckling J, Auyeung B, Happé F, Szatmari P, Baron-Cohen S, MRC AIMS Consortium (2019) Neural self-representation in autistic women and association with “compensatory camouflaging.” *Autism: The International Journal of Research and Practice* 23(5): 1210–1223. <https://doi.org/10.1177/1362361318807159>
- Lekić Barunčić K (2021) *Philosophical Perspectives on Autism*. Rijeka: University of Rijeka, Faculty of Humanities and Social Sciences.
- Lind S, Bowler D (2010) Episodic memory and episodic future thinking in adults with autism. *Journal of Abnormal Psychology* 119: 896–905. <https://doi.org/10.1037/a0020631>
- Lopez B R, Lincoln A J, Ozonoff S, Lai Z (2005) Examining the relationship between executive functions and restricted, repetitive symptoms of autistic disorder. *Journal of Autism and Developmental Disorders* 35(4): 445–460. <https://doi.org/10.1007/s10803-005-5035-x>
- Malatesti L, Jurjako M, Meynen G (2020) The insanity defence without mental illness? Some considerations. *International Journal of Law and Psychiatry* 71: 101571. <https://doi.org/10.1016/j.ijlp.2020.101571>
- Meyerding J (2014) Thoughts on finding myself differently brained. *Autonomy, the Critical Journal of Interdisciplinary Autism Studies* 1(3), North America, 1, Jul. 2014. Retrieved March 2, 2022. <http://www.larry-arnold.net/Autonomy/index.php/autonomy/article/view/AR9>
- Monsell S (2003) Task switching. *Trends in Cognitive Sciences* 7(3): 134–140. [https://doi.org/10.1016/s1364-6613\(03\)00028-7](https://doi.org/10.1016/s1364-6613(03)00028-7)
- Moore M S (2016) The Neuroscience of Volitional Excuse. In: *Philosophical Foundations of Law and Neuroscience*, eds Patterson, D. Pardo, M. S., 179–230. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198743095.003.0010>
- Murrie D, Warren J, Kristiansson M, Dietz P (2002) Asperger's syndrome in forensic settings. *International Journal of Forensic Mental Health*, 1, 59–70. <https://doi.org/10.1080/14999013.2002.10471161>
- Nelson R H (2021) A critique of the neurodiversity view. *Journal of Applied Philosophy* 38: 335–347. <https://doi.org/10.1111/japp.12470>
- Office of the Child Advocate State of Connecticut (2014) *Report of the Office of the Child Advocate*, p. 9, section 36. Retrieved March 2, 2022. <https://web.archive.org/web/20141126233253/http://www.ct.gov/oca/lib/oca/sandyhook11212014.pdf>
- O'Sullivan O P (2018) Autism spectrum disorder and criminal responsibility: Historical perspectives, clinical challenges and broader considerations within the criminal justice system. *Irish Journal of*

- Psychological Medicine* 35(4): 333–339. <https://doi.org/10.1017/ipm.2017.13>
- Ortega F (2009) The cerebral subject and the challenge of neurodiversity. *BioSocieties* 4(4): 425–445. <https://doi.org/10.1017/S1745855209990287>
- Otero T M, Barker L A (2014) The frontal lobes and executive functioning. In *Handbook of Executive Functioning*, eds S. Goldstein & J. Naglieri, 29–44. Springer. https://doi.org/10.1007/978-1-4614-8106-5_3
- Ozonoff S, Cook I, Coon H, Dawson G, Joseph R M, Klin A, McMahon W M, Minshew N, Munson J A, Pennington B F, Rogers S J, Spence M A, Tager-Flusberg H, Volkmar F R, Wrathall D (2004) Performance on Cambridge Neuropsychological Test Automated Battery subtests sensitive to frontal lobe function in people with autistic disorder: Evidence from the Collaborative Programs of Excellence in Autism network. *Journal of Autism and Developmental Disorders* 34(2): 139–150. <https://doi.org/10.1023/b:jadd.0000022605.81989.cc>
- Petrolini V (2020) Too much or too little? Disorders of agency on a spectrum. *European Journal of Analytic Philosophy* 16 (2): 79–99. <https://doi.org/10.31820/ejap.16.2.4>
- Reed P, McCarthy J (2012). Cross-modal attention-switching is impaired in autism spectrum disorders. *Journal of Autism and Developmental Disorders* 42(6): 947–953. <https://doi.org/10.1007/s10803-011-1324-8>
- Richman K A (2018) Autism and moral responsibility: Executive function, reasons responsiveness, and reasons blockage. *Neuroethics* 11(1): 23–33. <https://doi.org/10.1007/s12152-017-9341-8>
- Russell J, Jarrold C (1999) Memory for actions in children with autism: Self versus other. *Cognitive Neuropsychiatry* 4(4): 303–331.
- Silva J A, Leong G B, Ferrari M M (2004) A neuropsychiatric developmental model of serial homicidal behavior. *Behavioral Sciences & the Law* 22(6): 787–799. <https://doi.org/10.1002/bsl.620>
- Simons J S, Henson R N A, Gilbert S J, Fletcher P C (2008) Separable forms of reality monitoring supported by anterior prefrontal cortex. *Journal of Cognitive Neuroscience* 20(3): 447–457. <https://doi.org/10.1162/jocn.2008.20036>
- Sokhadze E, Baruth J, El-Baz A, Horrell T, Sokhadze G, Carroll T, Tasman A, Sears L, Casanova M F (2010) Impaired error monitoring and correction function in autism. *Journal of Neurotherapy* 14(2): 79–95. <https://doi.org/10.1080/10874201003771561>
- Stoyanov D S (2018) Psychiatry and neurolaw: An essay on the mind-brain problem and legal proof. *Balkan Journal of Philosophy* 10(1): 27–36. <https://doi.org/10.5840/bjp20181014>
- Stegenga J (2021) Medicalization of sexual desire. *European Journal of Analytic Philosophy*, 17(2) (SI5): 5–32. <https://doi.org/10.31820/ejap.17.3.4>
- Stokes M, Newton N (2004) Autism spectrum disorders and stalking. *Autism: The International Journal of Research and Practice* 8(3): 337–339.
- Šendula Jengić V, Grozdanić V, Hodak J (2022) Psychopaths – A “tough nut” of forensic psychiatry practice in the Republic of Croatia. In *Psychopathy: Its Uses, Validity and Status*, eds L. Malatesti, J. McMillan, P. Šustar, 59–78. Springer International Publishing. https://doi.org/10.1007/978-3-030-82454-9_5
- Volkmar F R, Loftin R, Westphal A, Woodbury-Smith M (Eds) (2021) *Handbook of autism spectrum disorder and the law*. Springer. <https://doi.org/10.1007/978-3-030-70913-6>
- Wakefield J C, Wasserman D, Conrad J A (2020) Neurodiversity, autism, and psychiatric disability: The harmful dysfunction perspective. In *The Oxford Handbook of Philosophy and Disability*, eds A. Cureton, D. T. Wasserman, 499–521. Oxford: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190622879.013.29>
- Zalla T, Daprati E, Sav A-M, Chaste P, Nico D, Leboyer M (2010) Memory for self-performed actions in individuals with Asperger syndrome. *PLoS ONE* 5(10): e13370. <https://doi.org/10.1371/journal.pone.0013370>

“DIE ERHEBUNG DES MENSCHENGEISTES ZU GOTT”. EVIDENCE AND PROOF OF GOD’S EXISTENCE, ACCORDING TO HEGEL

Dragoș Popescu

The Philosophy and Psychology Institute
Romanian Academy of Sciences
Bucharest, Romania
e-mail: dragos.gp@gmail.com

Abstract:

The paper examines the meaning of a Hegelian expression: “the ascension of the human spirit to God”, which was formulated in the philosopher’s 1829 summer course dedicated to the proofs of God’s existence.

We argue that the Hegelian formula describes a double movement: the first one refers to the departure of thinking from the Phenomenon and its arrival to the Ideal, and the second one describes the opposite movement, in which thinking crosses the barrier between the Ideal and the Phenomenon.

Key words: Proofs of the God’s existence, Phenomenon, Ideal, Necessity, Accident.

During the summer semester of 1829, Hegel taught a course that was complementary to his usual course of Logic. It was intended to analyze the evidence of God’s existence. The content of the course, kept in the teacher’s notebooks and divided into sixteen lessons, was included later in the posthumous volume *Vorlesungen über Philosophie der Religion* (*Lectures on the Philosophy of Religion*), as a kind of culmination of it¹. The subject of the

course, proper to the old metaphysics, was no longer at the forefront in an age in which it was commonly believed that metaphysics itself belonged to the past.

However, the lack of *philosophical* interest in the subject did not indicate a *total* lack of interest in it. In other words, if metaphysics had failed to convincingly guarantee the highest object of thought — and had therefore been set aside as a result of this failure — guarantees could still be received from other sources. Assessing these with the tools of philosophy is what Hegel proposes to us in his course.

What we propose in the present paper is not to review the philosopher’s evaluation — initiated in a certain historical context, but not from a historical perspective — of the most specific products of thought (as the proofs of the existence of God were understood to be for a long period of time), but to draw from his exposition a guide to his conception of “evidence” (*Beweis*) and “proof” (*Beweisen*).

Meiner, Leipzig, 1930 (The English version we use in this paper can be found in: G.W.F. Hegel, *Lectures on the Philosophy of Religion*, vol. III, London, K. Paul, Trench, Trübner, & co., L^{TD}., 1895, p. 153-367. The translator was Ebenezer Brown Speirs; see also the Romanian version: G.W.F. Hegel, *Dovezile existenței lui Dumnezeu*, traducere și note de Dragoș Popescu, Editura Tracus Arte, București, 2020).

¹ The German edition we follow is: G.W.F. Hegel, *Vorlesungen über die Beweise vom Dasein Gottes*, neu herausgegeben von Georg Lasson, Verlag von Felix

The “evidence” and the “proof” guarantee the scientific character of the knowledge. As a result, their study is the prerogative of logic (as Hegel conceived it, of course). The “proofs of God’s existence” are the bonds that hold the separate realms of Faith and Thought together. This union of the fields of Faith and Thought is specific to the Christian religion, according to Hegel. For Christians (that is, for all those who are part of the spiritual configuration of the Christian religion), the union of Faith and the Thought is a profound *need*.

Therefore, according to the above, the study of the evidence for the existence of God is of logical interest but, *at the same time*, it also responds to a need felt in the sphere of Christian spirituality. So, logic is the means (and the instrument) by which, only through thinking, we can guarantee that we are really dealing with evidence, not with something else.

1. The ascension of the human spirit to God

The reason why, in the sphere of Christianity, there is an intense need to prove the existence of God is rooted in a belief formulated by Hegel as follows: *Erhebung des Menschengeistes zu Gott* (cf. *Vorlesungen über die Beweise vom Dasein Gottes*, ed. cit., p. 13), that is: *the ascension of the human spirit to God, or the elevation of the human spirit to God* (cf. *Lectures on the Philosophy of Religion*, ed. cit., p. 164). It turns out that it is of primary interest to explore the meaning of this expression in order to better understand it, since the connection of logic (or: science in general) with religion (in this case: with the Christian religion) depends on it.

Of course, “ascension” (*die Erhebung*) indicates the unfolding of a transformation, of a process, at the end of which something becomes something else. We are dealing with two moments: one of departure and another in which the target was reached. The third, intermediate moment (cf. the verb *vermitteln*, the noun *Vermittlung*) is that of the actual transformation. The point of departure is the “human spirit” (*Menscheng Geist*) and the point of arrival is “God” (*Gott*).

If the transformation in question is *not* considered, any examination of the evidence for God’s existence becomes useless, that is it does not (any longer) correspond to any need, but remains an additional, secondary, ornamental, conservation approach, which we can possibly give up. Historiographic research (whose methodological perfor-

mance — in terms of strength and thoroughness — should not be underestimated), finding all sorts of forgotten aspects, patiently restoring neglected details of the evidence of God’s existence, always tends toward such side approaches. The danger posed by excessive historical research is precisely this: it may push the process of the ascension of the human spirit to God — whence came the interest in the proofs of the existence of God — to one side, leaving this crucial concept to be increasingly overshadowed by side approaches (cf. *Vorlesungen über die Beweise vom Dasein Gottes*, p. 51).

In addition to the difficulties faced by historical research in examining the evidence of God’s existence, in the ascension of the human spirit to God there is another tendency: to consider such evidence in a climate dominated by the emotional side of the relationship between man and God, a side of the relationship that man’s relationship with God is bound to presuppose. The emotional relationship that people have with God is inevitable; it cannot and must not be hindered. But considering the ascension of the human spirit to God exclusively through the prism of emotion (by diminishing the role of thinking) pulverizes the force of the evidence of God’s existence, reducing it to a bundle of divergent feelings, which no longer prove with certainty but rather produce the opposite: uncertainty (see *Vorlesungen über die Beweise vom Dasein Gottes*, p. 25).

If we turn our attention to the “ascension of the human spirit to God”, then another aspect appears: that of the nature of the movement that takes place (unfolds) through it. Of course, it is not appropriate here to consider some kind of physical movement (transport, or lifting). And, at the same time, we do not consider the “human spirit” and “God” as two abstract, conventional entities, like the material points in physics, located on two convergent trajectories.

The “ascension” of the Spirit takes place, of course, also in the Spirit. The moment of departure and of the end are both in the Spirit, of the Spirit, the Spirit itself. The Spirit itself undergoes, therefore, the transformation from one moment to another, being Spirit in both: starting point and purpose. He transforms. There is no distance between the “human spirit” and “God” like that between the Earth and the Sun, but rather like that between a bud and the corresponding fruit.

In Hegel’s view, the study of this type of *transformation* belongs to a systematic philosophical

discipline (in the sense that it is part of the whole system of science): the phenomenology of the spirit. The *results* of the transformations, in turn, belong to the Philosophy of Spirit, the third sequence of the encyclopaedical system of philosophy (composed of: logic, philosophy of nature, philosophy of spirit).

The proofs of God’s existence, therefore, involve almost all fields of philosophical investigation as understood by Hegel: the phenomenology of the spirit, logic, the philosophy of the spirit, the philosophy of religion. However, the areas in which knowledge *opposes* its object are not taken into consideration.

This is because, in the case of knowledge that is opposite to the object, there is a discontinuity (barrier) between the “human spirit” and “God”. The *opposition* between knowledge and object initiates a different process between the object of knowledge and knowledge than the process of the ascension of the human spirit to God. The former is the process by which knowledge and object are identified, so that knowledge is no longer about the object, and the object ceases to be the *purpose* of knowledge, being equivalent, that is *identical*, to knowledge itself. Knowledge, in the case of a partial identification with its object, is finite (delimited by the purpose it has to accomplish, to achieve, and which constitutes its limit). And finite knowledge is improper in relation to God (infinite).

The difference between the process of ascending the human spirit to God and the process of identifying knowledge with the object explains why Hegel argues that the evidence for God’s existence is particularly of interest in the Christian religious context. Only in this context was the second process fully realized, only *here* was the identification of knowledge with its object able to complete the journey to its own purpose. Although each knowledge / object configuration has, in its turn, concerns about the divine, infinity, eternity, etc., these concerns are exhausted depending on how the opposition between knowledge and object has been resolved there.

If we take an example from the Hegelian *Vorlesungen über die Beweise vom Dasein Gottes*, ed. cit., p. 96:

“Dem unabwendbaren Verhängnis nachzugeben, dazu ermahnten die Weisen, besonders die Weisheit des tragischen Chors, und wir bewundern die Ruhe ihrer Heroen, mit der sie ungebeugten Geistes frei das Los entgegennehmen, welches das Schicksal ihnen beschied”;

or, from the English version:

“That Man should yield to inevitable destiny was the advice of the wise, and this was in particular the truth expressed by the tragic chorus, and we admire the repose of their heroes and the calmness with which they freely and undauntedly accept the lot which destiny has assigned to them”
(*Lectures on the Philosophy of Religion*, ed. cit., p. 277).

This expressive presentation of the attitude of the Prophet and the Hero of the Greek tragedy (chorus leader, he Prophet: Tiresias; Oedipus: the Hero) in the face of relentless Necessity as engendered by the Gods hides, under apparent peace and reconciliation, sorrow and defeat by Necessity, not freedom and fulfillment in Necessity. Here, the Necessity (ἀνάγκη) of the Greek tragedy is the discontinuity, the limit, the barrier that the Hero encounters in the knowledge / object configuration, without being able to overcome it. Necessity, as it appears in the tragic conception of the world, did not remain an insurmountable, absolute barrier in the way of the human spirit:

“die Notwendigkeit selbst hat ihre Wahrheit in der Freiheit”
(*Vorlesungen*, ed. cit., p. 87),

according to the English translation:

“Necessity itself, however, has its truth in freedom”
(*Lectures*, ed. cit., p. 237).

The barrier was overcome by the human spirit, Necessity itself undergoing a transformation as Necessity.

The tragedy, however, asserts itself in its own spiritual configuration — reached by the Greek people after a laborious development of what we may call “ascension to Necessity” — as a kind of proof. Still, it is evidence *only* in relation to the Greek conception of Necessity. *For us*, who are beyond the barrier of Necessity (as the Greeks understand it), it no longer constitutes evidence.

The necessary transition from the “human spirit” to “God”, as we have already said, is not hindered by Necessity, as in tragedy, but makes the connection of one term (One) with another (Other). The starting point of the connection is accidental

(because it is the one that will undergo the change, the transformation), and the end of the process is absolutely necessary in relation to the starting point.

The transition from accidental to necessary cannot be considered according to the causal scheme. The reason why the transition from One to the Other is not intelligible (understandable) as a causal relationship is given by the absolute character of the Other in its relation to the One:

“das absolut Notwendige als Wirkung anzunehmen, fällt jedoch wohl sogleich hinweg; dies Verhältnis widerspricht dem absolut Notwendigen”

(*Vorlesungen*, ed. cit., p. 101),

or, in English:

“[The latter category, however, according to which] the Absolutely-necessary is taken as an effect, can be at once discounted, since the relation it implies is in too direct contradiction to the characteristic with which we are dealing, namely, the Absolutely-necessary”

(*Lectures*, ed. cit., p. 282).

The proper resolution of the knowledge / object opposition cannot be satisfied with the rejection of the causal relationship. Between the human spirit and God, any connection is broken as long as a model of the connection cannot be conceived; but we are not dealing only with this. Indeed, the realm of the finite is separated from that of the infinite; between science and feeling (faith) incompatibility is established, and the communion of humans disappears.

The strict separation of the finite domain from that of the infinite is, according to Hegel, *dogmatism*. The basis of this dogmatism is logic (see *Vorlesungen*, ed. cit., p. 109)². Dogmatism autonomizes, by interrupting the connection between the One and the Other, the domains of each, which means that the barrier of Necessity reappears and is re-established between them. The realm of the finite initiates its own process, different from that of the ascension of the human spirit to God, a process which is strictly delimited within the Necessity, and the “ascension” envisaged by it acquires a specific form:

“Die Bestimmung der Notwendigkeit, wie ihr spekulativer Begriff sich uns expliziert hat, ist überhaupt der Standpunkt welcher Pantheismus genannt zu werden pflegt”

(*Vorlesungen*, ed. cit., p. 126);

or, following the English version:

“The determination of necessity as unfolded in the philosophical conception of it, is, speaking generally, the standpoint which we are in the habit of calling Pantheism”

(*Lectures*, ed. cit., p. 316).

This pantheism is religious and/or philosophical (metaphysical). Limiting ourselves to the philosophical conception (only before Hegel; because after Hegel other new forms appeared), it is of various kinds: Hindu, Spinozist, Neoplatonist, etc.; but all are reunited by the scheme of the process in which it develops: the deification of the finite, realized through the metaphysical abstractions that perform the process within the finite.

Separating the domain of the finite from that of the infinite is not the solution to the relationship between them. Instead of this dogmatic solution, something else must be put in place, another model that allows the connection of domains. This model is drawn by Hegel in the following words:

“Das Moment, auf das hauptsächlich in diesem Zusammenhange aufmerksam gemacht worden, ist, daß er ein *Übergang* ist, d.h. daß das, wovon ausgegangen worden ist, darin die bestimmung eines Negatives hat, als ein zufällige Sein, nur als Erscheinung ist, welches seine Wahrheit an dem absolut Notwendigen, dem wahrhaft Affirmativen desselben habe”

(*Vorlesungen*, ed. cit., p. 118).

In the English version we find the equivalent:

“The moment to which attention has mainly to be directed in reference to this connection, is the fact that it is a transition, that is to say, the point of departure has here the characteristic quality of something negative, has the character of contingent Being, of what is a phenomenon or an appearance only, which has its truth in the Absolutely-necessary, in the truly affirmative element in this latter”

(*Lectures*, ed. cit., p. 305).

² We note, therefore, that dogmatism does not mean the coercive acceptance of a set of previously formulated statements, but the predetermination of intelligible domains by means of logical procedures.

These words clearly state that One (accidental) and Other (absolutely necessary) are:

1. each negative of the other (One is non-Other / Other is non-One);
2. Phenomenon (One) and *An-sich* (Other);
3. Absolute Necessity as a statement (existential statements: One is (= exists) / Other is (= exists) are *simultaneously* true, each being false).

The three conditions can only be met if we see the finite / infinite ratio as an passage, not a barrier.

2. Determination, Mediation, Exhaustion

“The ascension of the human spirit to God” is a process whose description requires the assistance of a special vocabulary. Terminology (the special study of the terms used in the theoretical exposition) was not an occupation encouraged by Hegel. However, clarifications of certain terms of particular importance cannot be avoided without introducing the possibility of misinterpretations. We thus address a few components of the vocabulary of the spirit’s ascension.

Determination (as action), *Determination* (as the result of the determining action), *Destination* (as the purpose of the action to determine) are words corresponding to the family: *bestimmen*, *Bestimmtheit*, *Bestimmung* etc. Here, the determining action refers to intelligible entities, conceptual components. Conceptual entities are either determined or determinant.

The Greek equivalent of the determination (*Bestimmung*) is ὄρος, meaning border, limit, barrier, trace. Within the perimeter delimited by the Greek word, the determinative is limited by the determinant. *Determinatio*, the Latin equivalent of the Greek word, is similarly composed around the root of *terminus*, meaning border.

Mediation also refers to the act of mediating, and of its outcome. In German there is a verb and a noun: *vermitteln*, *Vermittlung*. The idea of middle is also present in German (*Mitte*). Mediation is the act by which a determination is made. The Greek variant is: μεσότης; the meaning of the Greek term, which is “Proportion” (“Ratio”), should not be overlooked in this discussion.

The medium cannot be absolute. It *mediates*, it is between a first entity (One — τὸ πρῶτον) and a last entity (Other — τὸ ἔσχατον); the mediator performs his mediation between them. Therefore, its absoluteness is the mediation itself, its effectiveness.

The equivalent of the word “effectiveness” is *Wirklichkeit*, which also means “reality”.

Exhaustion (*Consumption*) is the famous Hegelian *Aufhebung*, about which so much ink has been spilled. The German term was equated with “suppression” (less often: “sublimation”); the meaning of “preservation” is also noteworthy here, and indeed it became a kind of reflex to use it to describe a process whose subject suffers an incomplete disappearance during its replacement with another subject. But this interpretation is not quite correct. The Greek verb καταργέω (from biblical source) better renders *aufheben*’s intentions. The corresponding noun is καταργέσις (used, for example, by Saint Paul in *Romans*, 3, 31).

However, *determination*, *mediation* and *exhaustion* are brought together by the overarching concepts of “overcoming” or “transience” (as quasi-equivalents for German: *Übergehen*).

3. Form of the Proof

“Syllogism” is an old technical term defined by Aristotle (although it is also used by Plato); it could be translated as conclusion = the intellectual action or operation by which a relationship between two terms is sanctioned (or: consecrated, recognized) based on their relationship to a third term. The equivalent German words are: *schließen*, *Schluß*.

But syllogism is not only an *operation*, it is also a *scheme* according to which the operation is performed. Unlike an operation, which lasts only as long as it is performed, the scheme persists – it can be repeated, memorized, inscribed on some medium, and used as a model for performing future operations. The evidence is syllogistic. It has a syllogistic form. The first meaning taken here by the expression “syllogistic form” is that of *scheme*, although it does not eliminate the *operation* meaning.

In the thirteenth lesson in the *Lectures on the Evidence of God’s Existence*, we find an illustration of “syllogism” as follows:

Because the Being is accidental, the Being is absolutely necessary³.

We notice immediately that this is not a scheme derived from Aristotelian logic (a syllogism such as Barbara, Celarent, Ferio, etc.).

Apparently, we are dealing with a statement like:

³ See *Vorlesungen über die Beweise vom Dasein Gottes*, ed. cit., p. 106: “Weil das zufällige Sein ist, so ist das absolut notwendige Sein”.

If X, then Y.

But, more specifically, the statement could be made:

If “Being is accidental”, then “Being is absolutely necessary”,
according to the equivalent:

X = Being is accidental,

respectively:

Y = Being is absolutely necessary.

The new logic (of course, “new” compared to “old”) indicates that:

$$X \rightarrow Y,$$

which reads: “X implies Y”.

However, slipping into this simplistic interpretation must be avoided. X and Y are not simple statements, they are what are called *modal* statements. In these, the *modus* and the *dictum* are different.

X and Y have the same *dictum*, but differ in *modus*:

Dictum: X = Being is (exists).

Y = Being is (exists).

Modus: X = It is accidental that «Being is (exists)».

Y = It is necessary that «Being is (exists)».

Given the above, we can say that, in terms of the scheme, here we are dealing with a transition from *accidental* to *necessary* with reference to a certain statement: “Being is (exists)”.

Returning to symbolization:

ACCIDENTAL (Being is) $\rightarrow$ NECESSARY (Being is)

or, using the usual signs of the modality (and, by “accidental”: possible):

$$\Diamond (\text{Being is}) \rightarrow \Box (\text{Being is}),$$

in simpler terms:

$$\Diamond (\Box X) \rightarrow \Box (\Box X),$$

or, to indicate the same thing, since X and Y have the same *dictum*:

$$\Diamond (\Box Y) \rightarrow \Box (\Box Y).$$

These two formulas deserve attention.

From the point of view of modal logic, it is contradictory to say that what is possible is necessary (as in the formula above). The correct logical transition from possible to necessary is done according to the formula:

$$\Diamond (\Box X) \rightarrow \neg \Box \neg (\Box X),$$

respectively:

$$\neg \Diamond \neg (\Box X) \rightarrow \Box (\Box X),$$

that is:

If it is possible for the Being to be, then it is not necessary for the Being not to be

and

If it is not possible for the Being not to be, then it is necessary for the Being to be.

We remember, however, from what was said above, in the first paragraph, that One (accidental) and Other (absolutely necessary) are each the negation of the other, but also that One lies in the phenomenal plan, Other in the ideal plan. By avoiding the contradiction, we actually interrupt the transition between the two plans.

However, at this point the solution to the problem is found: if, at the level of *dictum*, in terms of Being, its negation is non-Being (Nothing), in terms of modality (*modus*), the negation of the phenomenon isn’t Nothing, but the Ideal. The negation of the Ideal, in turn, isn’t Nothing but the Phenomenon:

“Nichts widerspricht sich nicht, es ist identisch mit sich; es erfüllt daher den logischen Satz, daß etwas sich nicht widersprechen sollte, vollkommen, oder wenn dieser Satz so ausgesprochen wird: nichts soll sich widersprechen, so es nur ein Sollen, das keinen Erfolg hat, den Nichts tut das nicht, was er soll, es widerspricht sich nämlich nicht”
(*Vorlesungen*, ed. cit., p. 119);

English equivalent:

“Nothing is in no sense contradiction. Nothing does not contradict itself, it is identical with itself; it accordingly fulfils perfectly the conditions of the logical proposition that a thing should not contradict itself — or if this proposition is expressed thus, Nothing ought to contradict itself, this is an ought which has no result, for Nothing does not do what it ought, that is, it does not contradict itself”
(*Lectures*, ed. cit., p. 307).

From the Hegelian point of view, however, this also evokes a much-quoted, yet misunderstood formula:

“Was vernünftig ist, das ist wirklich;

Und was wirklich ist, das ist vernünftig”

(G.W.F. Hegel, *Grundlinien der Philosophie des Rechts*, in Hauptwerke in sechs Bänden, Band 5, Felix Meiner Verlag, Hamburg, 1999, p. 14),

or:

“What is rational is actual;
and what is actual is rational”

(G.W.F. Hegel, *Elements of the Philosophy of Right*, edited by Allen N. Wood, translated by H.B. Nisbet, Cambridge University Press, 2003, p. 20).

4. Evidence, Proof

Then, we return to evidence and proof.

What is clear now is that Hegel proposes a kind of model for understanding the evidence that is somewhat different from the traditional one. Traditionally, the evidence (or arguments) for God’s existence has often been systematized, commented upon, and analyzed. The discussions about each argument individually, and about all of them together, make up a huge, inexhaustible literature. Rather than engage with this literature at any length, we will limit ourselves to transcribing one of many lists that enumerate these arguments⁴:

1. The Cosmological argument
 - a) from the contingency of the world
 - b) from the acknowledging of the movement
 - c) from the acknowledging of life on Earth,
2. The Teleological argument,
3. The Moral argument
 - a) the existence and function of consciousness
 - b) the notions of Good and Evil, Moral and Im-moral
 - c) the need for a reward based on our own acts,
4. The Ontological argument,
5. The Historical argument,
6. The Psychological argument,
7. The Noetic argument,
8. The argument *ab utili*,

...

None of the once well-known arguments just evoked (along with their variants, which each have numerous subvariants) uses the proof mechanism we

have presented; in fact, each argument of the list seems to be only a piece of that larger mechanism. Not one of these is comprehensive to the extent that it could replace the entire list. Each is an autonomous product of thinking, and the fact that they are all separate arguments, rather than elements of a more complete scheme, appears to be a circumstance that is not favorable for every argument.

However, if we consider them together, all the arguments about the existence of God are in fact based on a certain notion (i.e.: The World, The Purpose, The Moral Conscience, *consensus gentium*, The Soul, The Being, The Truth, The Utility), whose structure has been generalized unto absolutization. Absolutized, the structure of the notion from which we start appears as the sought after proof. However, any absolute notion is the notion itself raised to the status of Idea. The Idea of The World, The Idea of The Purpose, of Moral Conscience, etc. – each is related to its corresponding notion, highlighting itself as *a priori* in relation to it. Objections to the arguments concerning the *existence* of God denounced exactly this process, which is a process of *thinking*. But “the ascension of the human spirit to God” is not *only* a process of thinking. As it unfolds, this process accomplishes what is to be demonstrated, so that at the end of the process, we do not have a simple transition from one notion to its corresponding Idea, but the *realization* of the process. The proof, here, is the Reality.

We must not restrict the notion of “proof” (and, of course, the evidence finding activity) to its Hegelian *Erhebung des Menschen Geistes zu Gott*. In Hegel’s exposition, we can in fact recognize several forms of “proof”. They are briefly presented below:

Scheme 1. The process flows on the Phenomenon plan, according to the scheme:

Being → Idea

in which the process takes place at the level of the same plan, which makes mediation unnecessary. This process is not, in fact, a syllogism. *Medius terminus* is missing. We are dealing with an ascent of thought to the ideal (συναγωγή), known during antiquity.

Scheme 2. The process takes place on the Ideal level. Its scheme is:

Idea → Being

⁴ We transcribe the enumeration found in the volume: Ioan Mihălcescu, *Noțiuni de filosofie religioasă*, fifth edition, “Cugetarea”, Bucharest, 1936, p. 44 sq. (A textbook for high school students, in Romanian, that briefly presents each of the arguments listed for teaching purposes).

and it shows us that a syllogistic scheme does not apply here. Rather, the Idea is distributed to the Phenomenon plan and becomes representation. It is the reverse of the above — διαίρεσις.

Scheme 3. The process reaches only the Ideas plan, but returns, and is completed, in the Phenomenon realm; described by the diagram:

Being → Idea → Being

In this case, thinking projects Being to the Idea plan. However, the mediation that takes place at this level is a simple identification between the Idea and the Phenomenon, not a true mediation. Mediation, in its true, Hegelian sense, is not a loss (dissolution) of the individuality of the mediated and mediating terms, a dissolution of them in one of the mediated plans, together One and Other. A prestigious historical illustration of a process of this type is given by Eleatism. It establishes Being at the Idea level, but there it blocks the whole process of thinking.

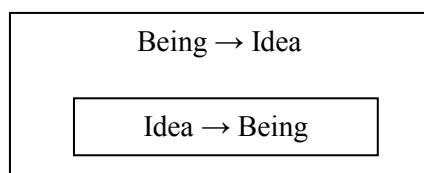
Scheme 4. The process divides the domains. The scheme is:



The scheme shows us that no mediation is possible. The two areas are strictly separate; in the absence of the *terminus medius* we cannot take into account any syllogistic structure.

The Enlightenment strongly illustrates this type of scheme. For the representative thinkers of this movement, separation of emotion and knowledge was also clearly stated. The need for mediation disappears completely in this variant process.

Scheme 5. The Hegelian scheme. It looks like this:



We immediately notice that, in this model, the *terminus medius* is itself a process (ideal). Even if this process is ideal, it takes place in the plan of the Phenomenon, which ensures the Reality of the whole process.

The Hegelian schema is, in fact, the only one in which we are dealing with a *terminus medius*. It is the only one in which the syllogistic process is not only initiated (as in Scheme 3), but can also be carried out in accordance with the traditional (Aristotelian) scheme⁵. It is the only scheme in which we can deal with *evidence* of God's existence.

References

Hegel G W F (1930) *Vorlesungen über die Beweise vom Dasein Gottes*, neu herausgegeben von Georg Lasson, Verlag von Felix Meiner, Leipzig.

Hegel G W F (1895) *Lectures on the Philosophy of Religion, together with a Work on the Proofs of the Existence of God* translated from the Second German Edition by the Rev. E.B. Speirs B.D. and J. Burdon Sanderson, the Translation edited by the Rev. E.B. Speirs B.D., in three volumes, vol. III, London, K. Paul, Trench, Trübner, & co., L^{TD}.

Hegel G W F (2020) *Dovezile existenței lui Dumnezeu*, traducere și note de Dragoș Popescu, Editura Tracus Arte, Bucharest.

Hegel G W F (1999) *Grundlinien der Philosophie des Rechts*, in Hauptwerke in sechs Bänden, Band 5, Felix Meiner Verlag, Hamburg.

Hegel G W F (2003) *Elements of the Philosophy of Right*, edited by Allen N. Wood, translated by H.B. Nisbet, Cambridge University Press.

Mihălcescu I (1936) *Noțiuni de filosofia religiei*, fifth edition, „Cugetarea”, Bucharest.

Popescu D (2016) *Termenul mediu în logica clasică și în logica speculativă*, în *Probleme de logică*, vol. XIX, Editura Academiei Române, Bucharest.

⁵ See our study *Termenul mediu în logica clasică și în logica speculativă*, în *Probleme de logică*, vol. XIX, Editura Academiei Române, București, 2016.

UNDERSTANDING AND IMAGINATION. A KANTIAN INTERPRETATION OF HEGEL'S "INVERTED WORLD"

Vlad Bilevsky

Romanian Academy Library
Calea Victoriei no. 125
010071 Bucharest, Romania
e-mail: vlad.bilevsky@biblacad.ro

Abstract:

In this article, I discuss an interpretation of Hegel's concept of the "Inverted World" (*verkehrte Welt*), which is present in the final part of the chapter on Force and Understanding in the *Phenomenology of Spirit*. Other than my own reading of the chapter, I also summarize the three most important interpretations of the *verkehrte Welt* from the last century: those of Jean Hyppolite, Hans-Georg Gadamer and Joseph Flay. I have chosen these three due to the typology of interpretation within them: the first one adopts a Christian reading of Hegel, the second a Hellenistic phenomenological interpretation, and the last one, which is closest to the interpretation I also propose, a reading consistent with the thought of Immanuel Kant. The article ends with a possible interpretation of Hegel's *verkehrte Welt* through Kant's "Schematism of the Pure Concepts of Understanding" from his *Critique of Pure Reason*, where I argue that Kant's fundamental faculty of imagination and its object, schemata, play the same role in Kant's system that Hegel's *verkehrte Welt* does in the development of natural consciousness.

Key words: Hegel, understanding, imagination, inverted world, natural consciousness.

The discussion of the "Inverted World" (*verkehrte Welt*) is one of utmost importance for a clear understanding – or experience, in this case – of Hegel's *Phenomenology of Spirit*. It is more so in this situation than in others due to its crucial position

in the book: it serves as an ending or transition point between the natural or intentional consciousness (*natürliche Bewusstsein*) and self-consciousness (*Selbstbewusstsein*). This metamorphosis which takes place within consciousness is one which also changes the consciousness's view of the world and, implicitly, of itself. Natural consciousness is one which sees itself as *different* from its object and is a *reactive* entity. In its movement from *sensation* to *perception* and then to *understanding*, consciousness ends a cycle. At the end of this cycle – and of every three part cycle, in fact – consciousness changes its ontological perspective from an "intellectualistic", intentional, dualist view to one in which consciousness identifies itself with its object and learns that in its attempt to know the world, it could only gain said knowledge insofar as it knows *itself*. The *verkehrte Welt* ends the discussion of natural consciousness in the third chapter, "Force and the Understanding", and is the final experience consciousness has before turning from a spectator of the world into an actor. Or, in Hegel's words: "*The necessary advance* from the previous shapes of consciousness for which their truth was a Thing, an 'other' than themselves, expresses just this, that not only is consciousness of a thing possible only for a self-consciousness, but that self-consciousness alone is the truth of those

shapes”¹. Despite its importance in the structure of the *Phenomenology of Spirit*, as Flay also claims, “it has remained for the most part one of the most obscure passages in the *Phenomenology*, a state of affairs which is usually attributed to the inherent absurdity of the position outlined in this passage”².

In this paper, I will attempt to demonstrate that Hegel's *verkehrte Welt* is not the “absurd, topsy-turvy” world that Flay and others suggest, but rather serves the purpose of being the phenomenological link between the phenomenon and its intellectual counterpart, lawfulness. In the course of my argumentation I will try to showcase an interpretation of the *verkehrte Welt* as a phenomenological correlative to Kant's Schematism of Pure Concepts of the Understanding, where he employs the transcendental to ground his view on the mediatory function of the *imagination*. Finally, I will juxtapose Hegel's corresponding passages alongside Kant's to illustrate a staggering similarity in ideas, which might show that *this* chapter of Kant's first *Critique* was what Hegel had in mind when writing about the *verkehrte Welt*. If we understand the “inversion” this way, the obscurity of the passage will suddenly become clear as we simply arrive at the analogy of the two philosophical systems. Thus, we will realize that what Kant thinks imagination does for the Understanding and sensibility, as the productive faculty of *schemata*, is what Hegel's *verkehrte Welt* does for the same counterparts. Unfortunately, Kant did not further expand upon the idea of the development of consciousness into self-consciousness despite reaching, with this hypothesis, the same conclusion as Hegel: *infinity*. In the end, in the second edition of the *Critique of Pure Reason* from 1787, Kant would also rewrite the imagination's functions and ultimately ended up modifying the chapter that Heidegger deemed the most important of the entire *Critique*³. We might add that if Kant had followed the development of his system with imagination as he originally conceived it, as the “common root” of sensibility and Understanding, he might have also arrived at Heidegger's conclusion that ultimately the *Critique*

of *Pure Reason* “has nothing to do with ‘the theory of knowledge’”⁴.

Before examining our main theme, I would like to point out the three main interpretations of Hegel's *verkehrte Welt* which have been generally used since the second half of the twentieth century until now: those of Jean Hyppolite, Hans-Georg Gadamer and Joseph C. Flay. My attempt to interpret Hegel's chapter using the analogy with Kant's “Schematism of Pure Concepts of the Understanding” does not, by any means, contradict these interpretations. We might say that the *Phenomenology of Spirit* holds, as it should, a different meaning to each individual reader, and thus, the science of the consciousness's experience. If we give a different interpretation than the three aforementioned, it goes merely to show that a reader can have a different experience with the passage than theirs, and can coexist in harmony, as the current interpretation only widens the variety of such endeavours.

Of the three, Hyppolite is the first to discuss the *verkehrte Welt* at great length. His interpretation is centered on a Christian view of the “two worlds”: “We might, in order to understand it [the experience of the inversion of the world – n. V. B.], refer less to science or the schellingian polarity and more to the dialectic of the evangels, which oppose the apparent world to the true world”⁵. The analogy here is made between the *Reich der Gesetze*, the kingdom of laws, and the *Reich des Himmels*, the kingdom of the skies. Following Hegel's examples, Hyppolite shows that “in the same way, in the evangel the one who is honored in this world is shunned in the other, that which seems to be power is weakness, the hidden simplicity of the heart is in itself superior to the apparent virtue”⁶. In the end though, Hyppolite also arrives at quite a Kantian conclusion: “The inverted world is not then to be searched for in a *different world*, it is present in this world here which is at the same time itself and its difference, which is known in its phenomenal integrality as ‘absolute concept’ or infinity”⁷. While in Christianity the *Reich des Himmels* is a different world than the phenomenal one, Hyppolite doesn't follow through with his analogy until the end. In fact, he arrives at the conclusion that the two worlds hold the same difference which

¹G. W. F. Hegel, *Phenomenology of Spirit*, Oxford University Press, 1977, translated by A. V. Miller, p. 102. The references to Hegel's *Phenomenology* throughout the article will correspond to this edition.

²Joseph C. Flay, *Hegel's "Inverted World"*, *The Review of Metaphysics*, vol XXIII, no. 4, June 1970, p. 662

³Martin Heidegger, *Kant et le problème de la métaphysique*, Editions Gallimard, Paris, 1953, translation by Alphonse de Waelhens and Walter Biemel, p. 186–187. Along with the transcendental deduction.

⁴*Ibidem*, p. 76–77.

⁵Jean Hyppolite, *Genèse et structure de la Phénoménologie de l'Esprit de Hegel*, Editions Montaigne, Paris, 1946, p. 132.

⁶*Ibidem*, p. 133.

⁷*Ibidem*, p. 134.

can be found within the distinction of the transcendental and the immanent. And, as we know, the transcendental "world" is not in a different realm than the phenomenon and can be said, in Hyppolite's own words, to be "at the same time itself and its difference".

The next thesis I will discuss is Hans-Georg Gadamer's. In this case, the author was formed in a phenomenological environment and, as such, we expect him to have a more Heideggerian approach. And while his ideas go on to develop a Platonic interpretation of the *verkehrte Welt*, we cannot ignore a passage which we find fundamental for our interpretation, and less for Gadamer's: "one cannot understand the beginning of the *Phenomenology* without keeping the Kantian philosophy directly in mind"⁸. Gadamer's interpretation would go on without any reference to Kant, but we can assume that he meant for it to apply to other passages of the natural consciousness rather than the *verkehrte Welt*. For this particular concept, Gadamer says: "We interpret this [the immediate elevation of the world of perception to universality – n. V. B.] to mean the *ascensus* of the Platonic archetypes, the ascent to a *noetic* world of perduring ideas"⁹. In the case of the *verkehrte Welt*, it is "a world which contains in itself the 'arché kineseos', and which really is the true world, is a reversal of the Platonic world which was supposed to be the negation of movement and change"¹⁰. For Gadamer, the *verkehrte Welt* is a type of Platonic *eidos* which holds the same relationship to the phenomenon as Plato's Ideas. Thus, in a Platonic expression, the Inverted World is "a world in which everything moves itself, because it contains in itself the origin of change"¹¹. His only addition to Hegel's text is that he does not consider the two supersensible worlds as *worlds*, because "in truth, it is not so much a question of the opposition of true worlds. Rather the 'true supersensible world' has both parts in itself: it opposes itself to its counterpole, and thereby alludes to it"¹².

The interpretation is truthful to Hegel's passage if you see the *verkehrte Welt* through the lens of the principle of change and alteration, which Hegel himself makes use of in the same chapter. It is this passage which is crucial to Gadamer's interpretation:

"With this [the inverted world – n. V. B.], the inner world is completed as appearance. For the first supersensible world was only the *immediate* raising of the perceived world into the universal element; it had its necessary counterpart in this perceived world which still retained *for itself the principle of change and alteration*. The first kingdom of laws lacked that principle, but obtains it as an inverted world"¹³. And to this passage, Gadamer adds: "it is the rational intention of reality on the level of the Inverted World to discard the empty universality of such hypotheses and laws, which are to a great extent overreached by reality itself. That which is rational and concrete is reality determined by the principle of alteration. Abstractions are again and again overthrown, for new ones to take their place"¹⁴. It is at this point that Gadamer searches for the Platonic dialectic in Hegel's thinking. And this is also where his interpretation ends, at the Greek dialectic of the *eidos* and *tode ti*, the individual.

While Hyppolite chose a Christian interpretation for Hegel's *verkehrte Welt* and Gadamer a phenomenological one, both agreed – the first one only in spirit, the other one openly – that Kant was essential to this chapter of Hegel's *Phenomenology*. Even so, neither of the two adopted a Kantian approach in this matter. The one to do so – and the last one we discuss before offering our own answer to the debate – is Joseph Flay.

Flay wrote his interpretation around the same time Gadamer did¹⁵. Despite that, their answers are polar opposites: while Gadamer retains a Hellenistic approach to Hegel's text, Flay is the first to view the *verkehrte Welt* with Kant and Leibniz as his guides. His perspective on the Inverted World is inspired by Kant's concept of the transcendental, which helps him understand Hegel's argumentation in a Kantian way – meaning that he posits absurd hypotheses – and this is why he calls the Inverted World "topsy-turvy", which he then considers absurd and from which he derives the actual thesis he wants to prove. If anything, Kant's argumentation is closer to a geometrical demonstration than a philosophical argumentation, and Flay follows this schema: "it is at this point [the principle of change and alteration, the same starting point for Gadamer – n. V. B.] where we reach a contradictory state of affairs and the discussion of the 'inverted world' arises. The law of appearance itself seemingly destroys the sought-after

⁸H. G. Gadamer, *The Inverted World, The Review of Metaphysics*, vol. XXVIII, no. 3, March 1975, p. 401.

⁹*Ibidem*, p. 411–412.

¹⁰*Ibidem*, p. 412.

¹¹*Ibidem*, p. 412.

¹²*Ibidem*, p. 418.

¹³G. W. F. Hegel, *Phenomenology of Spirit*, p. 96–97.

¹⁴H. G. Gadamer, *The Inverted World*, p. 414–415.

¹⁵Gadamer wrote his in 1966 and Flay in 1970.

unity and stability which is to constitute lawfulness. Laws as well as the concept of law itself embrace an identity in difference and a difference in identity”¹⁶.

We must note that a Kantian lecture on Hegel's passages on the Inverted World might not be true to the intention of the book, which, after all, suggests that consciousness goes through a gleaming, sorrowful process by gradually exposing itself to greater risks and greater suffering, without considering its past experience as absurd. It is, of course, scientifically absurd if we disregard the great lessons that consciousness receives from its failures, but an honest reader of Hegel must keep in mind that he is reading a book about “the science of the consciousness's experience”, which is the phenomenology of *spirit*. A Kantian lecture on such a text would simply prove the “absurdity” of the hypotheses posited throughout the work and would find such an endeavour fruitless. But if we choose to read Hegel with Heidegger's motto in mind – *Wege, nicht werke*¹⁷ –, then we have to be careful not to exaggerate the Kantian methodology and, in an effort to transpose it in a different environment, try to model said environment after it. This is, in our opinion, what Flay does here when he calls the *verkehrte Welt* “absurd”. He explains his method as such: “The task to be accomplished is the same as Kant's discussed above. The difference is that it is carried out, not by directly insisting upon the difference as such, but by drawing absurd conclusions which follow if and when the distinction between the old metaphysics (the first supersensible) and the new position (the second supersensible) is not made. Without the distinction absurd contradictions follow and the ‘second supersensible world’ in fact only serves to make unintelligible the world of appearance which it purports to ground and make intelligible. Sour becomes sweet, North Pole becomes South Pole etc. (...) The law of the second supersensible domain, as here misinterpreted, is in fact lawlessness”¹⁸. A Kantian demonstration and the Hegelian phenomenology are substantially different. Firstly, nowhere in Hegel's text do we see a distinction between “the old metaphysics” and “the new position”. The Inverted World doesn't make “the tranquil kingdom of laws” unintelligible, as lawfulness only contradicts the transcendental in *spirit*, due to the “principle of change and alteration”. The physician, for instance, would easily be able to contradict this label as noth-

ing is more “intelligible” – in a Hegelian sense – than science. Even if the two worlds are “generically different”¹⁹, the transcendental doesn't make laws unintelligible; rather, it only serves to *reinforce* said laws and eliminate the arbitrary immediate elevation from earlier and replace it with the transcendental deduction. So while it is true, as Flay concludes, that “a backward glance at Kant's distinction between phenomena and noumena, and the discussion of the amphibolies of reflection would help to make Hegel's intention clear”²⁰, Hegel does not arrive at the transcendental through the same means Kant employs, and thus the “supersensible” isn't the Kantian *noumenon* in its fullest sense. Instead, as we will try to show, it holds a closer relationship to the Kantian *schemata* and the *imagination*. In any case, if we must admit that absurdity is indeed present in Hegel's text, then it is not just the *verkehrte Welt* that is absurd, but *every world* becomes absurd when related to another. The phenomenal world is absurd for the “tranquil kingdom of laws” in the same way that the latter is absurd for the *verkehrte Welt* and so on. Absurdity is, understood in this way, in no sense different from the Hegelian negativity. All worlds are each other's negative and together form the configuration of natural consciousness.

I suggest a different Kantian analogy than that of the *phenomenon* and *noumenon*. In this interpretation, the three Hegelian worlds (phenomenality, lawfulness and the *verkehrte Welt*) each correspond to one of Kant's fundamental faculties: sensibility, Understanding and imagination, or their objects: intuitions, categories and schemata. I shall assume that the reader is at least somewhat familiar with Kant's first *Critique* and not discuss the genesis or functions of the sensibility and understanding in the Kantian philosophy, just as I have not summarized the “play of forces” that gave birth to lawfulness in the Hegelian chapter. Instead, I will focus only on the Inverted World's correlation in the Kantian text with the productive imagination.

For Kant, imagination and schemata are the necessary common root of the other two faculties, just as the *verkehrte Welt* is the one that “holds together” the other worlds for Hegel. Kant's “Schematism of the Pure Concepts of Understanding” begins by highlighting the function of the *schemata*: “In all subsumptions of an object under a concept the representations of the former must be *homogeneous* with the latter, i. e. the concept must contain that which is

¹⁶Joseph C. Flay, *Hegel's "Inverted World"*, p. 662.

¹⁷*Paths, not just works*.

¹⁸*Ibidem*, p. 675–676.

¹⁹*Ibidem*, p. 675.

²⁰*Ibidem*, p. 678

represented in the object that is to be subsumed under it, for that is just what is meant by the expression 'an object is contained under a concept'"²¹. The nature of this activity has to be transcendental, and this is where "it is pretty clear that there must be a third thing, which must stand in homogeneity with the category on the one hand and the appearance on the other, and makes possible the application of the former to the latter. This mediating representation must be pure (without anything empirical) and yet *intellectual* on the one hand and *sensible* on the other. Such a representation is the *transcendental schema*"²². In this scenario, the transcendental schema fills the role that will later be replaced, in the last *Critique*, with *judgement*: "Hence an application of the category to appearances becomes possible by means of the transcendental time-determination which, as the schema of the concept of the understanding, mediates the subsumption of the latter under the former"²³. This way, the relation of the schemata to the categories mirrors the Hegelian opposition between the two supersensible worlds, lawfulness and the *verkehrte Welt*: "this representation of a general procedure of the imagination for providing a concept with its image is what I call the schema for this concept. (...) In fact, it is not images of objects but schemata that ground our pure sensible concepts"²⁴. The process of knowledge finally takes, with the addition of the schemata, a new shape: "even less does an object of experience or an image of it ever reach the empirical concept, rather the latter is always related immediately to the schema of imagination, as a rule for the determination of our intuition in accordance with a certain general concept"²⁵. And Kant concludes: "Thus the schemata of the concepts of pure understanding are the true and sole conditions for providing them with a relation to objects, thus with *significance*"²⁶.

The Kantian passages could be transposed, in a Hegelian lecture, in the *Phenomenology's* "Force and Understanding" chapter. Kant starts the argumentation once he admits that sensibility and the Understanding – or the Hegelian phenomenon and lawfulness – are insufficient by themselves to serve as a

basis for knowledge, requiring a *third* faculty. This is also Hegel's conclusion in the passage about the *immediate* elevation of the phenomenon to concept that the understanding effects, as previously discussed²⁷. Without the *verkehrte Welt*, we could say that Hegel – or the natural consciousness, in our case – would arrive at Kant's conclusion: "Without schemata, therefore, the categories are only functions of the understanding for concepts, but do not represent any object"²⁸. This is also Hegel's conclusion in the following passage, from the same chapter: "This realm of laws is indeed the truth for the Understanding, and that truth has its *content* in the law. At the same time, however, this realm is only the *initial* truth for the Understanding and does not fill out the world of appearance. In this the law is present, but is not the entire presence of appearance; with every change of circumstance the law has a different actuality"²⁹. Understanding, as lawfulness, cannot by itself, fill the entire presence of the phenomenon. Perception, as Hegel discussed in the previous chapter, is not reliable either. Thus, Understanding *needs* the *verkehrte Welt* just as Kant needed *schemata*. Only in this way can Hegel say that "the inner world is completed as appearance"³⁰.

The major difference comes after this point, when Kant ends up abandoning imagination and the crucial role of time as the inner sense, while Hegel goes on to show that the two supersensible worlds are, in fact, a unity: "Thus the supersensible world, which is the inverted world, has at the same time overarched the other world and has it within it; it is for itself the inverted world, i. e. the inversion of itself; it is itself and its opposite in one unity. Only thus is it difference as *inner* difference, or difference *in its own self*, or difference as an *infinity*"³¹. And this way, "law completes itself into an immanent necessity"³², and the *verkehrte Welt* has achieved its purpose.

The attempt to portray Hegel's Inverted World through the Kantian Schematism of Pure Concepts ends here, as Kant could not reach the conclusion of self-consciousness, when the Understanding realizes that all knowledge is, in fact, self knowledge. But

²¹Imm. Kant, *Critique of Pure Reason*, Cambridge University Press, 1998, translation by Paul Guyer and Allen W. Wood, p. 271.

²²*Ibidem*, p. 272.

²³*Ibidem*, p. 273.

²⁴*Ibidem*, p. 273.

²⁵*Ibidem*, p. 273.

²⁶*Ibidem*, p. 276.

²⁷G. W. F. Hegel, *Phenomenology of Spirit*, p. 96: "For the first supersensible world was only the *immediate* raising of the perceived world into the universal element".

²⁸Imm. Kant, *Critique of Pure Reason*, ed. cit., p. 277.

²⁹G. W. F. Hegel, *Phenomenology of Spirit*, p. 91.

³⁰*Ibidem*, p. 96.

³¹*Ibidem*, p. 99.

³²*Ibidem*, p. 99.

Kant might've shared, as I have tried to show, a part of Hegel's road, although the two systems parted at this point.

References

- Bübner R (1975) Kant, Transcendental Argument and the Problem of Deduction. *The Review of Metaphysics* XXVIII(3): 453–467.
- Dove K R (1970) Hegel's Phenomenological Method. *The Review of Metaphysics* XXIII(4): 615–641.
- Findlay J N (1958) Hegel. A Re-Examination. New York: MacMillan Company.
- Flay J C (1970) Hegel's "Inverted World". *The Review of Metaphysics* XXIII(4): 662–678.
- Gadamer H-G (1975) The Inverted World. *The Review of Metaphysics* XXVIII(3): 401–422.
- Glockner H (1958) Die europäische Philosophie von den Anfängen bis zur Gegenwart, 3. Aufl., Reclam.
- Hegel G W F (1977) *Phenomenology of Spirit* (translated by A. V. Miller). Oxford: Oxford University Press.
- Heidegger M (1953) *Kant et le problème de la métaphysique* (traduction par Alphonse de Waelhens et Walter Biemel). Paris: Editions Gallimard.
- Hyppolite J (1946) *Genèse et structure de la Phénoménologie de l'Esprit de Hegel*. Paris: Éditions Montaigne.
- Kant I (1998) *Critique of Pure Reason* (translated by Paul Guyer and Allen W. Wood) Cambridge: Cambridge University Press.
- Kojève A (1985) *Introduction à la lecture de Hegel*. Paris: Editions Gallimard.
- Pöggeler O (1966) *Die Komposition der Phänomenologie des Geistes, "Hegel-Studien"*, Beiheft 3, Bonn: H. Bouvier u. Co. Verlag.
- Pöggeler O (1976) *Georg Wilhelm Friedrich Hegel, "Grundprobleme der grossen Philosophen"*. hrsg. von J. Speck: Philosophie der Neuzeit II, Göttingen, Vandenhoeck & Ruprecht.
- Siep L (2000) *Der Weg der Phänomenologie des Geistes. Ein einführender Kommentar zu Hegels «Differenzschrift» und zur «Phänomenologie des Geistes»*. Frankfurt am Main: Suhrkamp Verlag.
- Verene D P (2007) *Hegel's Absolute: An Introduction to Reading the Phenomenology of Spirit*. New York: State University of New York Press.

QUANTUM MECHANICS, BEING AND COGNITION

Ivan Katzarski

Institute of Philosophy and Sociology,
Bulgarian Academy of Sciences,
Sofia, Bulgaria
email: pessoko@gmail.com

Abstract:

This article examines the epistemological views of key quantum physicists of the Copenhagen circle (Niels Bohr, Werner Heisenberg and Léon Rosenfeld). The discussion begins with a presentation of their conception of measurement, indeterminacy and complementarity, and goes on to focus on their views regarding the nature of being and knowledge. The author identifies the basic areas of consensus in the Copenhagen circle as well as the disagreements and disputes that arose between its members.

Three main points are argued: (1) The fundamental epistemological consensus in the group was that subject and object are inseparable; that the subject participates in the formation of images of reality, which are multiple and depend on how the subject is concretely constituted historically; the disagreements within the group did not arise with logical inevitability from the principles of quantum mechanics but largely stemmed from ideological clash in the Cold War context. (2) Being linked to specific configurations of the subjects, knowledge is always relative; but it does not follow that knowledge is illusory or inadequate; on the contrary, it is the pretension of attaining the ultimate nature of things “in themselves” that leads to devaluation of any concrete knowledge. (3) The world is something external, generally independent of our desires and will, but at the same time is constructed as concrete objects through the specific configuration of each concrete subject.

Key words: Bohr, Heisenberg, Rosenfeld, epistemology, dialectics, reality.

1. Introduction

In a previous publication (Katzarski 2022), I analyzed some paradoxes in quantum mechanics. Atoms and subatomic particles display properties unlike those of bodies and forces in the macro-world; these new properties include particle-wave duality, the absence of continuity in space and time, the impossibility of tracing the transformations occurring across the states of particles. The only laws that prove equally valid in the macro and micro world are those of conservation of energy, momentum and charge. The categories valid in the macro-world (causality, discontinuity and continuity, processuality, space and time) lose their validity in the world of microparticles.

This is the framework in which we examine the crisis of classical rationality and of the language that could adequately describe quantum phenomena. We argue that the mathematical language which successfully deals with the paradoxes of quantum physics is not, and cannot be, sufficient in itself, and that it inevitably engages in dialogue with other languages that are more concrete and closer to common human experience. Moreover, due precisely to the paradoxical nature of quantum mechanics, even its creators found it convenient to resort to images typical of magical-mythical mentality and poetry. The discontinuity of quantum phenomena is reminiscent of the magical transformations undergone by characters in myths and fairy tales.

Thus, science, which in its pursuit of absolute objectivity has progressively freed itself of anthropomorphic representations, has found within the depths of matter (as well as in the singularities of cosmic evolution) certain images analogous to the images and metamorphoses proper to an archaic mentality.

Under these circumstances, the question of the epistemological status of quantum mechanics naturally arises. How can the scientific findings of quantum mechanics be interpreted? We know about Einstein's doubts: he believed quantum theory was incomplete. But disputes as to its epistemological status were going on more or less openly even among the adherents of the Copenhagen current. The problem involved is so fundamental that it ultimately touches on how we understand nature and human knowledge in general. These questions are integral to our analysis here.

2. Measurement, indeterminacy, complementarity

Confronted with the paradoxical behavior of microparticles, proponents of quantum mechanics developed a methodological sensitivity regarding research procedures in physics; and measurement is at the basis of all these procedures, as reliable knowledge in this science is built upon the empirical basis established through measurement.

The adherents of the Copenhagen perspective on quantum mechanics, especially Niels Bohr, Werner Heisenberg and Léon Rosenfeld, had a strong focus on the methodological aspects of their work.

According to Rosenfeld, measurement can never be reduced to the registration of sense data. The process of measurement always involves a certain idealization that disregards some of the changes taking place in the structure of the measuring instruments and some of the deviations arising in the course of measurement. Moreover, accuracy in the procedure requires knowledge of the physical laws on which measurement is based. Hence, measurement always requires a considerable amount of knowledge (Rosenfeld 1979: 625).

In classical mechanics, measurement has two aspects: one related to energy and another related to space and time. It can be carried out simultaneously in its two aspects: the measurement of energy (momentum) of movement with the desired degree of accuracy is no obstacle to registering the spatial-temporal frame with equal accuracy. The measuring instruments in both cases are in principle mutually compatible, and only when their respective data are

taken together can a physical system be fully described (Rosenfeld 1979: 266-267).

The situation is radically different in quantum mechanics. Here, researchers face a fundamental difficulty: "In fact, any unambiguous use of the concepts of space and time refers to an experimental arrangement that involves a transfer of momentum and energy, uncontrollable in principle, to instruments such as measuring rods and synchronized clocks, which are required for fixing the reference frame. Conversely, an account of phenomena governed by conservation of momentum and energy involves in principle a renunciation of detailed space-time coordination. (...) Under these conditions, it is not surprising that phenomena observed with different experimental arrangements appear to be contradictory when it is attempted to combine them in a single picture" (Bohr 1956: 87).

In classical mechanics, the presence of measuring instruments does not influence the behavior of the measured object, i.e., the macroscopic body. To the contrary, "in the region of atomic physics (...) it is no longer possible sharply to distinguish between the autonomous behavior of a physical object and its inevitable interaction with other bodies serving as measuring instruments, the direct consideration of which is excluded by the very nature of the concept of observation in itself" (Bohr 1937: 290).

And thus, in the light of quantum objects, fixing the space-time framework necessarily involves some degree of indeterminacy with regard to the energy aspect. And vice versa: the fixing of the energy dimensions introduces a similar indeterminacy with regard to space and time. Werner Heisenberg has defined this situation as "a relation of indeterminacy" (Unbestimmtheitsrelationen) (Heisenberg 1931: 176-177).

The relations of indeterminacy correspond, of course, to the wave-particle duality of the microparticles, which cannot be expressed simultaneously through the dimensions of energy and space-time. However, this problem is of a fundamental, logical sort: how are the two aspects of quantum reality (which we can measure only separately, alternatingly) mutually related?

At the level of facts, the relations of indeterminacy are not a topic of dispute. The main dissension occurs in the interpretation of these facts. Several solutions emerge, which are mutually exclusive and give rise to heated dispute. Erwin Schrödinger's concept "tries to limit the scope of quantum theory to the sole wave aspect. To him the field governed by the

wave equation is the ultimate reality; the particle concept would just be a name for the element of discontinuity introduced by the 'quantisation' of the field" (Rosenfeld 1979: 472). Albert Einstein did not challenge the veracity of the empirically registered relations of indeterminacy, but he believed that the quantum theory based on them represents a transient conception that will be overcome by a future theory in the classical vein, in which there will be no place for indeterminacy and stochasticity (Einstein 1934: 168-169; 1940: 492). The solution proposed by David Bohm is generally similar. According to him, "a description of the phenomena entirely in the spirit of corpuscular mechanics would supply a deterministic substratum for quantum mechanics, whereas the de Broglie wave would only occur as an auxiliary concept; its function would be to represent in some fashion an average residual interaction of the atomic system under consideration with the macroscopic bodies which fix the conditions of the phenomenon" (Rosenfeld, 1979: 475). Bohm in fact was developing an earlier idea of de Broglie's that the latter had abandoned.

However, the prevalent solution to be established in physics was different: it did not try to abolish indeterminacy by reducing one of the two aspects of quantum phenomena. The solution in question is found in the interpretation proposed by the Copenhagen group, with Bohr in the lead. With wonderful simplicity, he contended that the two aspects were equally important and mutually complementary. Bohr wrote: "The apparently incompatible sorts of information about the behavior of the object under examination which we get by different experimental arrangements can clearly not be brought into connection with each other in the usual way, but may, as equally essential for an exhaustive account of all experience, be regarded as 'complementary' to each other" (Bohr 1937: 291).

Far from concealing the fact, Bohr actually stressed that this principle was in contradiction with classical rationality. But according to him, this did not mean we had left the domain of logic and science: "Far from containing any mysticism foreign to the spirit of science, the notion of complementarity points to the logical conditions for description and comprehension of experience in atomic physics" (Bohr 1958: 172). He also stressed that the relations of indeterminacy could not be annulled or overcome by a future expansion of knowledge. Hence, the hopes for a return to classical rationality were in vain, for the issue was not of a concrete scientific nature, but related to

logic and epistemology. The return was just as impossible as it would be to refute Einstein's theory of space-time by returning to the idea of absolute space and time (Bohr 1937: 293-294).

3. Epistemological consensus and disputes

It would not be an exaggeration to say that quantum mechanics is a direct path to the core of philosophy, to the questions of the nature of being and knowledge. The situation is unprecedented: while the rise of physics is largely linked to this science's emancipation from philosophy – and even to its adoption of a haughty attitude toward the latter – the theory of relativity and quantum mechanics provoke philosophical reflection and debate in the fullest sense. The heated debate as to the nature of micro-objects represents, in fact, a clash between different philosophical visions of the world and man, nature and knowledge. Critics of quantum mechanics (like Einstein, Bohm and many others) defended the classical objectivism that postulates the existence of an external world to be cognized, "reflected" correctly, and thereby changed. The Copenhagen group aimed to overcome this view by assigning a role to the knowing and acting subject.

The two viewpoints are actually very familiar in philosophy. The first of these has the support of so-called "common sense" and the categories of classical scientific rationality. The latter viewpoint is in contradiction with "common sense" but has on its side the basic facts and arguments of 20th century domains of knowledge. That is why I think the philosophical views of the Copenhagen group are the more interesting. However, as we will see, the group's views present certain important intra-group differences along with their common elements.

Werner Heisenberg

In his schema of the structure of reality, Heisenberg distinguishes a "'most objective' domain of reality we can speak of without referring directly to the methods with which we learn of it"; this includes, firstly, that part of the world that is described in the terms of classical physics (Heisenberg 2019: 39). Here, "objectivity" means that the results do not depend on the applied instruments and methods. But this independence is possible only because the material systems "are sealed off from the external world" and their "temporal course is determined by their present state for all time" (Ibid.: 63). In the world of microparticles, such an "objectification" would be impossible because the system under study cannot be

separated from the instruments and methods used to study it. There is no “objective result” separable from and independent of the activity of the knowing subject (Ibid.: 63).

We find, in fact, that even in classical physics it is only exceptionally that the object of study can be fully isolated and have its future state calculated based on an initial state. The really interesting cases are when such isolation is impossible because all the factors on which the state of the system depends cannot be encompassed. The factors not taken into account lead to errors, which accumulate and make it impossible to predict the behavior of the system beyond a certain moment in time. Such an essential incompleteness of knowledge is a factor, for instance, in the case of weather forecasts or when calculating the trajectories of planets in the solar system (Ivanov 2012: 36-37).

Of course, the fundamental inexhaustibility of knowledge of classical objects is different from the impossibility of differentiating between the states of microparticles and the techniques used to measure those states. In the classical case, the degree of indeterminacy of the system can be reduced by encompassing more factors pertaining to the environment; in the latter case, the indeterminacy is fundamental and primary. In the world of quantum phenomena, a magnitude that characterizes the quantum state “in general denotes not a certain property of the system accessible to the senses but a specifically defined range of possibilities pertaining to all properties accessible to the senses. It is only through the act of observation that something objective is created from these possibilities. It follows that this act of observation and the intervention necessitated by observation are a decisive feature of quantum theory and its subject matter” (Heisenberg 2019: 56-57).

Ernst Fischer makes an interesting comment on this point: in his lecture ‘Language and Reality in Modern Physics’, Heisenberg “writes explicitly, that physicists got used to viewing electron orbits and similar terms not as a reality but a kind of ‘potentia’ in the sense of Aristotelian philosophy”. Fischer adds that “the Greeks put being and possibility side by side, and presumed that it would require a force to change the possible into reality. For this force they introduced the term ‘Energeia’, and hence energy, which plays a prominent role, not just in physics.” Thus, according to Heisenberg, quantum mechanics is “the first theory that considers side by side a ‘res extensa’ (reality) and a ‘res potentia’, as a realm of possibility (...) The world presents itself as ‘potentia’

from which, by observation, an actual reality is gained” (Fischer 2019: 138).

I should note that this associating of Heisenberg with Aristotle is interesting but misleading. Heisenberg’s interpretation was still quite close to the experimental data, while Aristotle had employed extremely abstract concepts that had nothing to do with experience. Concepts such as ‘virtual particles’ and ‘virtual orbits’ are elements of modern physical theory, while Aristotle’s notion of matter as a potentiality devoid of all determinacy and his ideas of active form and the cosmic first mover belong to the sphere of pure metaphysics. The direct translation of concepts from one sphere to the other is unfounded.

But let us go back to Heisenberg’s reasoning. He applies his thesis about the unpredictability of the subject not only to quantum mechanics. In his manuscript ‘Reality and Its Order’, Heisenberg views the other domains of reality (organic life, the conscious and the unconscious, religious life, social organization, etc.) in analogy with quantum mechanics. He grounds his analysis on a specific notion of reality. As the distance from the domain of classical physics increases, “the process which provides information about reality, itself increasingly forms part of the relevant complex of connections that constitute the relevant domain” (Heisenberg 2019: 38). Ultimately, “we must always bear in mind that the reality of which we can speak is never reality ‘per se’ but a perceived reality, even, in many cases, one we ourselves have shaped” (Ibid.: 38-39). And also, “The concepts ‘objective’ and ‘subjective’ designate two poles from which an order of reality can take its beginning. They also describe two sides of reality itself. Still, it would be a crude oversimplification to divide the world into an objective and a subjective reality. This merely black and white representation created much inflexibility in philosophy of the last few centuries” (Ibid.: 34). The next assertion is no less telling: “Ultimately, cognition is perhaps nothing other than an ordering process – not of something that would have already been accessed by our consciousness or perception but of something that requires this ordering process in order to become the actual content of consciousness or a process we are aware of” (Ibid.: 101).¹

¹ Here I am referring to Heisenberg’s general concept of reality. I will not comment on his characterization of the separate domains of being – this would require a special analysis that is unnecessary for the purpose of this discussion.

As mentioned, Ernst Fischer finds an existing kinship between Heisenberg's conception and the philosophy of Aristotle. Another author asserts that at the Göttingen seminars organized by Carl Friedrich von Weizsäcker, Heisenberg "was deeply committed" to the Kantian view (Mohanty, J. N. 1989: 384). In his writings, however, Heisenberg shows no attachment to Kant. On the contrary, the main motif there is a critique of Kantian apriorism, described as incompatible with modern physics and with modern science in general (see for instance: Heisenberg 1931: 175–176, 181–182; 2000: 51, 48, 49–50; 2019: 31).

It is true that both Kant and Heisenberg pointed out the irreducibility of knowledge to any pure objective external content. But there is a fundamental difference between them. For Kant, the knowing subject was absolute: his forms of sensibility and categories are unchanging, given once and for all. For Heisenberg, to the contrary, the subject is historically concrete; the practical and theoretical instruments he has at his disposal and applies change significantly over time, and as they change, quite different pictures of the world appear. 'Reality and Its Order' presents such differing pictures, which are dependent on basic factors such as organic evolution, phases of psychological development of the individual, the technological potential of society, the domination of some basic attitudes and the suppression of others.

Besides criticizing Kant's apriorism, in the 'thing in itself' Heisenberg sees nothing of importance for modern physics; he only mentions this term in passing (Heisenberg 2000: 51).

Heisenberg had set his heart on another philosophical tradition, that of Pythagoras and Plato. According to Heisenberg, the concept of "atom" in Leucippus and Democritus is the foundation of materialism; in modern physics, however, he sees this concept interpreted in a new way that distances us from materialism (Heisenberg 2000: 26); with regard to the elementary particles, "one sees that not even the quality of being (if that may be called a 'quality') belongs to what is described. It is a possibility for being or a tendency for being" (Ibid.: 34–35). The conclusion is clear: "The elementary particles in Plato's *Timaeus* [which are, of course, different from what Heisenberg and modern physics mean by them – I. K.] are finally not substance but mathematical forms. 'All things are numbers' is a sentence attributed to Pythagoras. (...) In modern quantum theory there can be no doubt that the elementary particles will finally also be mathematical forms, but of a much more complicated nature", than in Pythagoras and Plato (Ibid.: 35–36).

Heisenberg's philosophical views have one more basic aspect: his attitude to dialectics. In the texts he published during his lifetime, we find next to nothing about this question. In 'Physic and Philosophy' (first edition, 1958) he pointed out that "modern physics is in some way extremely near to the doctrines of Heraclitus", specifically to this Greek philosopher's concept of fire as "both matter and a moving force" (Ibid.: 29).

We find much more on this issue in Heisenberg's manuscript 'Reality and Its Order', dating from 1941–42 and published posthumously. Here the dialectical understanding underlies the author's general concept of reality. Heisenberg distinguishes between two basic aspects of reality: the 'static' and the 'dynamic'. The more rigorous a science is, the more its language is unambiguous, specialized and formal. It is a language capable of describing a limited domain of reality, and incapable of revealing connections with other domains or serving as a basis for a more integral understanding of the world. According to Heisenberg, mathematics and jurisprudence are especially vivid examples of such narrow conceptualization. As for the dynamic aspect of reality, it is attainable, on the contrary, thanks to the indeterminacy and ambiguity of language:

It is made possible by those infinitely complex associations among words. An idea expressed in this way is not meant to be as faithful a representation of reality as possible but to be the seed for further series of ideas. The issue is not the accuracy but the fruitfulness of concepts. (...) Here, a sentence can, generally speaking, not be 'right' or 'wrong.' But one may call a sentence 'true' when it fruitfully leads to an abundance of other ideas. The opposite of a 'right' sentence is a 'false' one. But the opposite of a 'true' sentence will often be another 'true' one. The most famous systematic formulation of this 'dynamic' representation of reality is Hegel's dialectics. The domain of 'static' thinking explains, since, after all, clarity is the goal of this form of thinking. The 'dynamic' domain interprets as it searches for the infinitely complex associations among other domains of reality to which we can point (Heisenberg 2019: 26–27).

He adds that the static and dynamic aspects are equally necessary but, taken separately, insufficient to describe reality: "Purely static thinking is in danger of deteriorating into form without content. Dynamic

thinking can become vague and incomprehensible” (Ibid.: 28). Heisenberg specifically underscores the importance of the dynamic aspect for describing new domains of reality: “The abyss between the already familiar conceptual systems and a new system may be leaped over by intuitive thinking but cannot be bridged by formal reasoning” (Ibid.: 28).

I could sum up Heisenberg’s philosophical position with three points. First, reality, or the world, is not something that exists independently of the cognizing and acting subject. This subject is part of reality and, through its inherent powers and abilities, participates in the formation of reality. It is not some abstract consciousness but a concrete subject formed through evolution and in history. It takes part in the construction of reality through its specific physical characteristics, sensory and mental abilities, technological potential, theoretical attitudes and approaches. The world is different for different living creatures, for the human individual in the consecutive phases of his development, for common sense vs. the scientific mind, for classical vs. modern physics, for people of different cultures, etc.

Secondly, Heisenberg is critical of materialism and shows a partiality for the traditions of Pythagoras and Plato, i.e., for the idealist trend. In my opinion, however, the interpretation of modern scientific facts, including the facts of quantum mechanics, does not oblige us to choose a definite philosophical viewpoint from among those which have been in irreconcilable conflict for many centuries. What modern physics defines as matter is not some concrete substance present to sensory perception, such as earth, water, fire or air, as some pre-Socratics believed; the matter of modern physics, according to Heisenberg, is only seemingly similar to the atoms of Leucippus and Democritus. Aristotle’s notion of matter as a passive substance lacking all concrete determinacy, as opposed to the active principle of form (the idea of the unmoved mover of the universe) likewise is without analogy in modern science. As for Descartes’s view on matter as a three-dimensional extension, it is in direct contradiction with modern representations of the micro-world. The fundamental correlation between mass and energy according to Einstein, as well as particle-wave duality, can be fitted within the frameworks of the classical doctrines of materialism, idealism or dualism, but only with some measure of artificiality.

Third, Heisenberg’s ideas regarding the static and dynamic aspects of reality, and their fundamental inseparability, are particularly valuable in philosophic

terms. Particularly significant in our time is the critique leveled by the mathematician and physicist Heisenberg against rigorously defined concepts and formalized systems. He indicates how important it is to preserve the elements of indeterminacy and dynamism of concepts – elements traditionally associated with dialectics in general, independent of dogmatic qualifications such as “materialist” or “idealist” dialectics.

Léon Rosenfeld

But the Copenhagen interpretation did not propose a single philosophical view of reality. Heisenberg’s conception had a no less interesting philosophical alternative.

Rosenfeld was part of the core Copenhagen group, which included, in addition to Bohr, also Pauli, Heisenberg and Jordan (Jacobsen 2007: 5). After 1930, and until Bohr’s death, Rosenfeld was the teacher’s closest associate. After World War II, Rosenfeld figured “as a particularly fierce and uncompromising defender of the Copenhagen Interpretation of quantum mechanics” (Ibid.: 5).

Like other members of the group, he considered the source of fundamental difficulty in measurement to be the microscopic size of the object of study in contrast to the measuring instruments that are macroscopic, i.e., composed of an enormous number of atoms. Rosenfeld stressed that these instruments could not be other than what they were, because they corresponded to the macroscopic nature of the human cognitive apparatus. He even believed that such must be the cognitive apparatus of any rational being capable of creating a theory of the micro-world (Rosenfeld 1965: 224-225).

But how does the process of measurement itself occur? The quantum object under study interacts with the macro-instrument, leaving permanent traces upon the latter:

How does such a permanent mark come about? Since it is of macroscopic character, it cannot be a direct result of the initial atomic interaction; the latter is rather of the nature of a triggering effect. It starts off within the measuring apparatus a macroscopic reaction, terminating when a state of stable equilibrium is reached, in which the desired characteristic mark has appeared. The triggering process only lasts for a time interval of the order of magnitude typical of individual atomic reactions; the ensuing macroscopic process necessitates a ‘relaxation time’ which, although it may be very short on

the macroscopic time scale, will in any case involve a large number of atomic reactions. This is the decisive point: the formation of the permanent mark is a process of ergodic character, entailing the wiping out of all structural details of the initial state of the total system, known as the 'reduction' of this state (Rosenfeld 1965: 225).

Clearly, this reduction "has nothing to do with the interaction between this system and the measuring apparatus: in fact, it is related to a process taking place in the latter apparatus after all interaction with the atomic system has ceased" (Ibid.: 230).

This analysis once again reveals the irreducibility of the subject (with all its physical and mental characteristics and its technological potential) within the domain of quantum mechanics. But based on this, Rosenfeld draws much broader, philosophically significant, conclusions. Cognition of nature is not a passive contemplation but a specific type of interaction with nature: "We are not merely contemplating the world, but acting upon it and thereby modifying its course. Accordingly, the scientific description of the phenomena is fundamentally concerned with the interaction of external agencies with the human observer; or, at least, in the narrower domain of physics, with material systems under the latter's control" (Rosenfeld 1979: 498). Thus, "the interaction between observer and phenomena appears as an inherent part" of the process of cognition (Ibid.: 521).

Rosenfeld applies this conclusion to the principle of complementarity between the space-time and the energy aspects of elementary particles. According to him, the question, "is this complementarity a property of the atoms themselves, or is it imposed by us on the atoms? (...) is a badly formulated question: the two partners – the atoms and ourselves – are inseparable" (Rosenfeld 1979: 557).

Man is part of reality. That is why his consciousness and knowledge cannot be somehow similar to a mirror, but are products of a specific interaction between the external environment and the processing of external stimuli, an interaction that can be defined as encoding. The sensory apparatus (receptors, nerves, cervical centers) transforms the external factors into mental symbols. But this is only the first stage of encoding and symbolization. Further on, the sensory information is subjected to encoding through the symbolism of language and the formation of concepts. The conceptual, mental encoding itself proceeds in various forms depending on the range of the initial encoded content and the level of abstractness

and logical connection between the concepts and the linguistic symbols. A number of assertions by Rosenfeld are in this vein (see for instance: Rosenfeld 1979: 333-334, 619-620).

He constantly points out that his understanding of man, being, mind and knowledge have nothing to do with subjectivism or subjective arbitrariness. Objectively, by the very nature of things, man is part of nature, and the knowledge of nature can in no way be independent of man as a subject; his active presence cannot be eliminated. "At first sight it would seem that one thus introduces in science a subjective element. But the fear that the objectivity and rationality (...) would be abandoned is entirely unfounded. True objectivity is simply the possibility of guaranteeing that the account of the phenomena will convey equivalent information to all observers, that it will consist of statements intelligible to all human beings" (Ibid.: 515-516).

These elements of Rosenfeld's conception are almost fully analogous to Heisenberg's views. One other similarity between the two is the positive attitude that both had towards dialectics. However, in the case of Heisenberg this attitude is expressed in a manuscript that was not published in his lifetime, while dialectics was a basic motif in all of Rosenfeld's works. According to the latter, through the concept of complementarity, Bohr had offered a dialectical view of nature: "The relation of complementarity is the first example of a precise dialectical scheme, whose formal structure has been accurately analysed by the logicians. This is a new fact whose importance could not be exaggerated" (Rosenfeld 1979: 481; see also: Ibid.: 262-263; Jacobsen 2007: 13).

We see that Heisenberg's and Rosenfeld's views on fundamental issues are astonishingly close and, on certain points, completely coincide. At the same time, however, their declared philosophical positions are diametrically opposed. Heisenberg asserts his sympathy for the line of Pythagoras and Plato, while Rosenfeld is a proponent of materialism: "[A]s far as quantum mechanics is concerned, I would say that it is impossible to understand it without assuming that there is an external world which is independent of what we think and which is the ultimate origin of all our ideas. In that sense I absolutely reject the suggestion of present-day positivists about the subjectivity of our statements" (Rosenfeld 1979: 501).

Not only was Rosenfeld a materialist, he was also a Marxist. In this respect, however, he proved unacceptable to the upholders of orthodox Marxism at

that time. He sharply criticized “mechanistic materialism”, under which he classified Lenin and the Soviet philosophers. Modern materialism, he asserted, could only be dialectical, and he saw this exemplified in the works of Marx and Engels. His antipathy towards Lenin was combined with sympathy for Mach (Rosenfeld 1979: 471–472, 682; Jacobsen 2007: 8–9, 32). As a staunch supporter of the Copenhagen interpretation of quantum mechanics, Rosenfeld sharply criticized the Soviet physicians who rejected it on ideological grounds (see for instance Rosenfeld 1979: 473–475).²

His criticism was justified. But in general, the philosophical battles over quantum mechanics are among the many examples of how ideas can fall victim to politics – in this case, to the politics of the Cold War, especially in its earliest and fiercest phase. And this applies not only to intellectuals in the Soviet Union and its satellite countries. Resistance to the Copenhagen interpretation was strongly motivated by ideology not only in the Eastern bloc but also in the West, especially in France (see: Cross 1991; Graham 1966).

Although the Copenhagen group was distinguished by its highly tolerant attitude, the differences between worldviews within it did not pass unnoticed. Most of the members (Born, Heisenberg, Pauli) criticized Rosenfeld for linking the principle of complementarity to materialism and Marxism. Despite these disagreements, Born, Heisenberg and Pauli remained close friends of Rosenfeld’s (Jacobsen 2007: 14, 28–29).

In my opinion, the philosophical-political disputes between the members of the Copenhagen group amounted in large part to a ‘comedy of errors’. Rosenfeld was not very convincing when he argued the principle of complementarity could be grounded in Engels’ ideas on nature and knowledge.³ His crit-

ics, however, could not boast of having a good enough knowledge of Marxism to criticize him in a well-argued way; what they often referred to was Stalin’s “abridged version” of Marxism. An example of such lack of understanding, of politicized philosophy, and of rather rude qualifications is to be seen in Max Born’s extensive discussion on modern physics and dialectical materialism contained in a letter to Rosenfeld (Freire and Lehner 2010). It is noteworthy that Bohr, the leader of the Copenhagen group, did not personally intervene in the disputes between Rosenfeld and the rest of the group members (Jacobsen 2007: 29–30). I believe this choice was a sign of maturity and wisdom on the part of a man standing above the ideological clashes of those times.⁴

Political détente (after the denunciation of Stalin) and the ideological sobering up of many Marxists (after the Soviet invasion of Hungary) had a positive effect on the quality of concepts and discussions. The case of Rosenfeld is very important in the theoretical aspect per se. His “Marxist vocabulary was considerably toned down in public beginning around 1960” (Jacobsen 2007: 34). But the change was not only in his way of expressing himself but also in the greater clarity and consistency of his concepts. In an article dating from 1970, he wrote, “[t]he conflict between materialism and idealism, under various forms, has played its historical part. If Heisenberg had been content to state that it has no longer any part to play in modern science, there would be no quarrel between us” (Rosenfeld 1979: 687). More than being just an observer, man interacts with that which exists around him. “It is clear that this relationship between the material universe and the human mind transcends the antiquated opposition of materialism and idealism, but could rather be described as a synthesis of these mutually exclusive points of view, in the sense of complementarity” (Ibid.: 330).

Here we see a more logically consistent and undogmatic standpoint, in the framework of which “objectivity” is no longer taken in the absolute sense it has in traditional materialism. The requirement of “objectivity” is no longer thought of as a full or maximal elimination of the subject’s influence in the course of cognition. The only sensible usage of the

² It would be hard to find a more severe judgment than this: “It is a cruel irony of fate that physicists who believe themselves the guardians of the Marxist tradition should be blind to... a momentous dialectical development and oppose to it the most outworn metaphysical preconceptions. Having failed in the attempt to uphold determinism in quantum theory they are now trying to argue that one ought to consider a kind of hybrid causality, partaking in some confused way of both determinism and chance... They are thus driven to a frantic chase of determinism into an infinite sequence of deeper and deeper levels of reality: a purely imaginative construction, as ingenious and misguided as those of medieval scholasticism” (Cross 1991: 744–746).

³ He was partially right as concerns dialectics in the works of Engels. But his views on causality, necessity and

chance were entirely in the spirit of classical science and as such were quite alien to the principles of quantum mechanics.

⁴ I touch on the ideological squabbles between physicists only in passing, because this article intends to emphasize the positive influence of modern physics on the classical philosophical problems of being and knowledge.

term, according to Rosenfeld, is to be had when it signifies “the consensus of all observers placed in the same conditions” (Ibid.: 550).

To sum up, Rosenfeld shared and defended the central thesis of the Copenhagen interpretation of the inseparability of subject and object, which complement each other to form what we perceive as nature or the external world. Specific to Rosenfeld was his attempt to prove that this understanding was a new form of materialism that overcame the outdated, metaphysical version. After the mid-1950s, however, Rosenfeld was inclined to reject the opposition between materialism and idealism, and to share the emphasis on the common epistemological basis of the two, which was typical for the Copenhagen circle.

4. Relativity of knowledge and truth

In the Copenhagen perspective on quantum mechanics, cognition is thought of as a specific type of interaction between subject and object. The result depends not only on external impact but also on the quality of the instruments the subject has at its disposal. Moreover, the result is neither an ideal mirror of reality nor a more or less dim mirror image; the subject plays a positive role because it takes part in the construction of the object.

Given this, the products of cognition are inevitably relative, insofar as they depend on the various configurations of the subjects. Is this relativity a vice, and is it rooted in some deficiency of the Copenhagen interpretation?

It is paradoxical that the deficiency and illusoriness of cognition can be argued on the basis of a contrary interpretation of quantum mechanics. Interesting in this respect are some reflections by Angel Stefanov, who shares Einstein’s view of the incompleteness of quantum mechanics which has a methodological basis (Stefanov 2020: 11–12). Stefanov discovers an inconsistency in the positions of the proponents of the Copenhagen interpretation: they do not offer a “structural model of the main object” of study in quantum mechanics – that is of the “quantum object per se, and despite this they insist that the theory is complete in the theoretical aspect” (Ibid.: 12). In the spirit of Einstein, he qualifies the Copenhagen interpretation as incomplete, because its proponents “have taken the phenomenalist nature of the principle of complementarity to be a cognitive revelation” (Ibid.: 12–13).

This understanding implies that genuine cognition reveals its objects as they are “in themselves”. Thus, “naïve realism” is misleading: grass is not

green, the stone is neither hard nor cold. These properties are not inherent to the objects themselves but are the product of a complex processing of the external impacts on our sensory apparatus and brain (Stefanov 2019: 6–7). It turns out that our daily experience is, not merely in exceptional cases, but essentially and radically, illusory as it does not reveal the “nature of things” “in themselves” (Stefanov 2019: 6–7).

Continuing this reasoning, Stefanov additionally sees the same “cognitive inadequacy” as applying to some basic representations of classical science deriving from everyday experience. Foremost in this regard are the classical concepts of space and time; as Einstein has demonstrated, space is not a “universal receptacle for physical objects”, nor are space and time independent essences. The inadequate representation of space and time, however, has taken hold because it proved useful for “the biological survival of primates millions of years ago” (Ibid.: 7–8).

Finally, according to modern physical theory, “space-time is also not a fundamental essence of nature, because it is said to be an emergence of more fundamental underlying objects, which are in themselves non-spatial and non-temporal” (Ibid.: 8–9).

And thus, proceeding from the statement that a given knowledge does not reflect the nature of things in themselves, the conclusion is drawn that knowledge is inadequate, illusory. Evidently the aim is to achieve a kind of knowledge that does express the nature of things. But there is no guarantee that any new knowledge is capable of achieving this ideal. Such is the conclusion to be drawn from the whole history of human cognition. And how could it be otherwise, given that cognition is always the work of a concretely defined and structured subject that inevitably sets its stamp on the obtained cognitive content? The cognizer is part of the world, and his abilities and instruments participate in the formation of images of being.

All knowledge is relative insofar as it is not, and cannot be, the work of some absolute subject. But this does not necessarily mean that it is inadequate or illusory. I believe it would be rather artificial to assert that the feelings of suffering from polar cold or equatorial heat are inadequate because they do not express some imaginary “nature of things” in themselves. Classical physics and, in general, classical science do not lose their cognitive and technological significance with the expansion of the horizon of science through the theory of relativity and quantum mechanics. The adequacy of knowledge must always be considered in

relation to a subject possessing a specific profile. If under precisely defined conditions of observation or experimentation we obtain recurring results, this signifies that the knowledge in question is adequate.

We see that the idea of some absolute sort of knowledge leads to the rejection of already acquired knowledge. And that knowledge is conceivable only as concrete, as related to a concrete subject, and therefore as relative.

What philosophers call “naïve realism” helps us survive and exist. Moreover, it reveals a basic truth: outside of us, and within us for that matter, something exists that opposes us, that we cannot control, that does not depend on our desires and will. Yet we could control it to some degree, providing our behavior is adequate to this independent reality, i.e., if our behavior is beneficial to our existence. What is the “deepest essence” of this opposed-to-us being is a matter for further philosophical speculation.

I find no convincing arguments that we should reject this understanding of the world. But it is only part of the truth about the world. For, although this reality is independent of our desires and will, its image is formed by qualities and instruments that human beings possess as specifically formed subjects of cognition. Thus, the external acquires a kind of objectness that is specific to each subject. For all living creatures, the external is invariably something that restricts them and stands in opposition to them, but in all else the external is different for each creature. And although in the sensory aspect reality is universally configured for Homo sapiens, its mechanism and logic of functioning are radically or significantly different for the magic-mythological mind vs. the scientific mind, for the small child vs. the adult, for the adherents of different religious and philosophical doctrines, for 19th century physicists vs. scientists who have adopted the theory of relativity or quantum mechanics, for members of different social groups, etc. The world is simultaneously external, in principle independent of our desires and will, and also constituted in its concrete objectness by the specific configuration of each concrete subject.

References

- Bohr N (1937) Causality and Complementarity. *Philosophy of Science* 4(3): 289–298.
- Bohr N (1956) Mathematics and Natural Philosophy. *The Scientific Monthly* 82(2), 85–88.
- Bohr N (1958) On Atoms and Human Knowledge. *Daedalus* 87(2): 164–175.
- Cross A (1991) The Crisis in Physics: Dialectical Materialism and Quantum Theory. *Social Studies of Science* 21(4): 735–759.
- Einstein A (1934) On the Method of Theoretical Physics. *Philosophy of Science* 1(2): 163–169.
- Einstein A (1940) Considerations concerning the Fundamentals of Theoretical Physics. *Science, New Series* 91(2369): 487–492.
- Fischer E P (2019) Commentary on Werner Heisenberg’s ‘Reality and Its Order’. In: . *Reality and Its Order* Heisenberg, Werner:. Springer Nature Switzerland AG.
- Freire O, Lehner C (2010) Dialectical Materialism and Modern Physics. An Unpublished Text by Max Born. *Notes and Records of the Royal Society of London* 64(2): 155–162.
- Graham L R (1966) Quantum Mechanics and Dialectical Materialism. *Slavic Review* 25(3): 381–410.
- Heisenberg W (1931) Kausalgesetz und Quantenmechanik. *Erkenntnis* 2: 172–182.
- Heisenberg, Werner. (2000). *Physics and Philosophy: The Revolution in Modern Science*. Penguin Classics.
- Heisenberg W (2019) *Reality and Its Order*. Springer Nature Switzerland AG.
- Ivanov M G (2012) How To Understand Quantum Mechanics? Moskva-Izhevsk: NIZ “Regularnaya i haoticheskaya dinamika”.
- Jacobsen A S (2007) Léon Rosenfeld's Marxist Defense of Complementarity. *Historical Studies in the Physical and Biological Sciences* 37(Supplement): 3–34
- Katzarski I (2022) Quanta, Rationality and Archaic Images. *NotaBene* 55. http://notabene-bg.org/uploads/articles_pdf/notabene-bg_org-read1267.pdf.
- Mohanty J N (1989) Idealism and the Quantum Mechanics. *History of Philosophy Quarterly* 6(4): 381–391.
- Rosenfeld L (1965) The Measuring Process in Quantum Mechanics. *Progress of Theoretical Physics*. Supplement, E65: 222–231.
- Rosenfeld L (1979) *Selected Papers*. Dordrecht/Boston/London: D. Reidel Publishing Company.
- Stefanov A (2019) ‘Owls Are Not What They Are’: An Unsought Revelation of Modern Science. *Filosofski Alternativi* 6: 5–13.
- Stefanov A (2020) Quantum-Logical and Constructive Approaches to the Microworld. *Filosofski Alternativi* 6: 5–14.

BOOTSTRAP'S MONADOLOGY. SYMMETRY AND MIRRORING CONNECTIONS BETWEEN CHEW'S BOOTSTRAP THEORY AND LEIBNIZ'S MONADOLOGY

Ramona Ardelean

Department of Socio-Human Science,
Politehnica University of Bucharest, România
email: ramona.ardelean.m@gmail.com

Abstract:

The scientific paradigm which I rely upon in the framework of this article is quantum mechanics, whose “cognitive revolution” consisted of replacing the classical principle of separability with the principle of nonseparability or global intercorrelation. According to this intercorrelation, highlighted at the subatomic level, the part cannot be separated from the whole, because every part has a global and instantaneous connection with the whole universe. For this reason the foundation of the world cannot be the part (elementary particles), but the whole, which is therefore logically and ontologically prior to the part, i.e., self-consistent. Consequently, the principle of global intercorrelation elucidates and validates some of the oldest philosophical problems and intuitions about the unity or self-consistency of the world. An example in this sense is the bootstrap theory of American physicist Geoffrey Chew, which presents such striking similarities to the metaphysical system of Leibniz's Monadology that the two intertwine and mirror each other, like twin souls, to the extent that it could be stated that if Chew's bootstrap theory represents the explanatory *physical* level of Leibniz's metaphysics, then, analogously, Leibniz's Monadology represents the explanatory *metaphysical* level of Chew's physics.

Key words: quantum bootstrap, monadology, symmetry, mirroring, part-whole intercorrelation, reciprocal causality, mutual self-consistency .

1. The quantum mechanics revolution: non-separability or global intercorrelation

The most significant “cognitive revolution” of the twentieth century is that of quantum mechanics, which affirms that the world is a coherent and indivisible whole that cannot be analysed in terms of separable¹ constituents of reality, as claimed by classical atomist-Newtonian physics. On the contrary, the world appears as a web or dynamic “interweaving” of events and interactions in which *every part is globally and instantaneously connected to the entire universe*. As such, an action which

¹ It is useful in this sense to reiterate that the classic principle of knowledge, respectively *the principle of separability between subject and object*, is based on the following presuppositions: (1) *the belief in an external/objective reality*, separated from the subject/internal, subjective reality, from which follows (2) *the status of the objective observer*, which excludes the subjective factor, as well as (3) the classic *distinctions*: subject-object, consciousness-matter (*res cogitans-res extensa*), part-whole, circumscribed to *the logic of parts*, in which the part is considered to precede the whole, and is therefore primordial, thus involving (4) *the substantialist/ fundamentalist conception* of the fundamental constituents, from which results (5) *determinism* – the interpreting of causality through mechanisms which are strictly *local*, i.e. localizable or determinable in space-time (See R. Ardelean 2020, 63).

affects a part is in fact affecting the whole, and vice versa, such that the part is influenced by all events in the whole (Capra 1975, 290). Considering that the part cannot be separated from the whole, it follows that the fundament of the world is no longer the part, but the whole, which is thus logically and ontologically prior to the part. Physicist David Bohm confirms this fact most eloquently:

One is led to a new notion of unbroken wholeness which denies the classical idea of analysability of the world into separately and independently existing parts...We have reversed the usual classical notion that the independent 'elementary parts' of the world are the fundamental reality, and that the various systems are merely particular contingent forms and arrangements of these parts. Rather, we say that inseparable quantum interconnectedness of the whole universe is the fundamental reality (Bohm 1975, 10).

The principle of quantum nonseparability thus denies the separability of the world in favour of its unity or nonseparability. Heisenberg affirms that "the world appears as a complicated tissue of events, in which connections of different kinds alternate or overlap or combine and thereby determine the texture of the whole" (Heisenberg 1963, 96).

The transition from separability to nonseparability has been facilitated by observations at a deeper level of matter – atomic and subatomic – which indicate that the principle of separability, which appeared plausible at the macrophysical level, is no longer valid at the microphysical or quantum level of reality. Since at this level there are no material objects with mass and weight, the particles under observation no longer appear as *objects with a substantialist character*, but as processes, events, or "probabilities of interconnection", which describe nonseparability or global intercorrelation.

The notion of global intercorrelation, initially brought forward by Niels Bohr and Werner Heisenberg, representatives of the Copenhagen school at the end of the third decade of the twentieth century, represents beyond divergent interpretations the element of convergence or quantum theory quintessence, in which *the behaviour of the part is*

always determined by its global connections, that is by its connections with the whole. This connection or universal interconnectivity denies any fundamental constituent, any elementary particle or isolated and separate entity, considered an abstraction or idealisation without a correspondent in reality. Niels Bohr demonstrates in this sense that isolated material particles are nothing more than abstractions, as one is only able to observe and define the properties of particles through their interactions with other particles or systems (See Bohr 1934, 57), or more precisely with the whole of which they are part. This means that, contrary to classical physics, in which predictions about hidden variables depend on a *local* mechanism, since the probability of producing events is determined by a local system in which the properties and behaviour of parts determine the behaviour of the whole, in quantum physics, the hidden variable is attributed to a *global, non-local* mechanism in which the whole "determines" (in the indeterminist sense of Heisenberg's uncertainty) *the behaviour of the part*, since the probability of an event being produced is determined by the dynamic of a global/non-local system in which every part has a *spontaneous and instantaneous* (non-local) *connection* with the whole universe, based on *mutual global intercorrelation*.

2. From global intercorrelation to self-consistency – Chew's bootstrap theory

Probing the subatomic world has revealed that all phenomena in the universe are interconnected, being parts which have spontaneous and global connections with the whole. Quantum mechanics thus abolishes at the subatomic/microphysical level the classical distinctions of the subject-object, observer-observed, consciousness-matter, part-whole type and replaces the principle of separability with the principle of nonseparability or global intercorrelation (See Bohm 2002, XI-XIV). The shift in paradigm facilitated by quantum mechanics thus gravitates around the understanding that the world does not consist of isolated, separate entities, but of inseparable parts of an indivisible, coherent and consistent whole.

By revealing the fundamental dimension of universal interconnection, quantum mechanics is closer than ever to the metaphysical view, both Eastern and Western, of the unity of things. The principle of quantum nonseparability (which is not at all philosophically neutral) *validates by means of scientific instruments the oldest metaphysical intuitions about the unity of the world*, thus confirming the statement of physicist Abner Shimony that, “experimental results have come to elucidate philosophical problems” (Shimony 1988, 78).

The *bootstrap theory*, proposed by American physicist Geoffrey Chew² is one such experimental result of quantum physics, supporting the elucidation and validation of some metaphysical intuitions. This theory presents, as we shall see, such evident similarities to the metaphysical system of Leibniz's Monadology that one can speak of a relation of *symmetry* or *mirroring* between Chew's physical theory and Leibniz's metaphysical theory. Thus, that which was explained by Leibniz at a metaphysical level is so clearly reflected in that which Chew explained at a physical level that one can state that *Leibniz's metaphysics “mirrors” Chew's physics* and vice versa. It is precisely by virtue of this symmetry that American philosopher George Gale points out, in a famous article titled *Chew's Monadology*, that Chew's bootstrap theory “can be taken to be the physical explanatory level correspondent to the Monadology” (Gale 1974, 348).

Chew's Bootstrap Theory

Since the 1960s, the notion of quantum intercorrelation, according to which the universe cannot be reduced to entities, particles or elementary fields, has led to the idea that the universe could manifest an even more subtle intercorrelation, that of self-consistency, through which all its components are at the same time “consistent both with one another and with themselves” (Capra 1975, 286).

This idea of self-consistency was formulated by Geoffrey Chew on two levels of understanding: a restricted one, limited strictly to scientific aspects, as well as a broader, philosophical one, un-

limited by experimental rigors and scientific formalism. Therefore, in the broadest sense bootstrap philosophy³ embraces the idea, “much older than particle physics, that nature is as it is because this is the only possible nature consistent with itself” (Chew 1968, 762). In a restricted sense, the scientific bootstrap elaborated in the framework of the mathematical formalism of the S-matrix⁴ considers only the particles which experience “strong interactions”⁵, i.e., hadrons, the aim of the hadronic bootstrap being “to derive all properties of hadrons and their interactions from the requirement of self-consistency” (Capra 1975, 294).

Rejecting the traditional, atomistic-Newtonian view based on elementary entities or particles, the criterion of self-consistency, suggested by Chew, states, on the contrary, that the universe is a dynamic, indivisible and coherent whole in which the properties of the part are connected with the properties of the whole, *none of which is fundamental* but which describe the property of the completeness of the mutual relations that determine the structure of the whole network. “In the new worldview, the universe is seen as a dynamic web of interrelated events. None of the properties of any part of this web is fundamental; they all follow from the properties of the other parts, and overall consistency of their mutual interrelations determines the structure of the entire web” (Capra 1975, 286).

Consequently, the great importance of S-matrix theory resides in the understanding of subatomic particles, not as objects, but as processes, events or relations wherein no elementary particles are assumed to be fundamental, but all particles and interactions arise in a self-consistent manner in keeping with the requirements of analyticity (causality), unitarity (conservation of probability), and crossing symmetry (See Chew 1968, 763-4).

If the traditional point of view considers that the bound-states were made up of elementary particles, Chew's more far-reaching vision showed, on

³ Philosophy of connection and intercorrelation

⁴ The letter S comes from the word “scattering”, referring therefore to the scattering matrix.

⁵ There are four categories of interactions: strong, electromagnetic, weak and gravitational. As strong interactions are hundreds of times more powerful than the other three, neglecting these, states Chew, seems a reasonable (scientific) approximation (See Chew 1968, 763).

² Geoffrey F. Chew (1924–2019) was a prominent scientist, teacher, and mentor. He is best known for developing the bootstrap concept as embodied in S-matrix theory.

the contrary, that there would be a new type of theory which describes the interactions of bound-states which have no point-like constituents at all. This approach was called “nuclear democracy”, since it avoided singling out certain particles as elementary (See Chew 1968, 764.).

Given that the bootstrap theory is a theory of networks of subatomic particles in which the properties of each particle derive from its relationships to the others, the only problem, according to Chew, is that of ever being able to build a complete S-matrix according to the bootstrap hypothesis, which is to say a matrix which would “contain” all the properties of the hadrons. “A bootstrapped S-matrix, Chew states, contains an infinite number of poles, and no single one can be completely understood without an understanding of all the others” (Chew 1968, 764-5), which is obviously impossible. Subsequently, to advance the bootstrap idea further automatically implies leaving the scientific field, which would require one to accept a number of hidden variables without any demonstration. By delimiting the scientific bootstrap theory from the bootstrap philosophy, Chew reveals that “in the broad sense, the bootstrap idea, although fascinating and useful, is unscientific. Science, as we know it, requires a language based on some unquestioned framework. Semantically, therefore, an attempt to explain *all* concepts can hardly be called ‘scientific’” (Chew 1968, 762-3).

As it is highly unlikely, from a scientific point of view, to ever be able to build a complete and consistent S-matrix, the only solution glimpsed by Chew was that of certain partial models, which would cover a group of limited interactions and contain a few hidden, unexplained parameters. These would then be explained within the framework of another model. This would consequently result in a “mosaic” of intercorrelated models whose number of unexplained parameters would continuously decrease and in which the term “bootstrap” would not fit an individual model, but an ensemble of mutually consistent models in which none would be fundamental in character. It is interesting to observe that S-matrix theory itself is such a “different” model – a partial model, resulting from the model of the quantum point of view (based on global intercorrelation) and the dynamic-relativist model. Accordingly, it is part of that “mo-

saic” of correlated models in a sense that is “democratic” rather than hierarchical or “aristocratic”. Any physicist, says Chew, who is capable of analysing a number of theoretical models without considering one more important than the other is automatically an adherent of the bootstrap point of view (Chew 1970, 27). Here, Chew is probably referring to an attitude of epistemological fair-play specific to the bootstrap that is also called nuclear democracy, which rejects as unfounded (at a subatomic level) the traditional point of view that privileges elementary parts, since this attitude is one that Chew considers “aristocratic” or hierarchical (See Chew 1968, 764).

The bootstrap theory, based on denying fundamental entities and understanding subatomic particles not as object, but as processes of movement, change and transformation, consequently reveals a structure, an (S-)matrix of the world, which is more likely based on the *verb* than on the noun (i.e., those “substances” or elementary entities arbitrarily postulated from the outside). Thus, if a theory’s degree of approximation to reality is determined by those “fundamental constants”, unexplained, but postulated, then the difference between modern theories and classical ones resides in the *temporary postulation* (Chew 1968, 763) of these constants, which need to be explained and integrated as the degree of precision and knowledge of the newest theory increases, thus aspiring to the ideal situation “where the theory does not contain any unexplained ‘fundamental’ constants, and where all its ‘laws’ follow from the requirement of overall self-consistency” (Capra 1975, 288). It should be noted, however, that it is possible such a situation will never be reached.

3. Chew’s Monadology and Leibniz’s Bootstrap: crossing symmetry and mirroring

Even though Chew’s bootstrap model follows the line of conceptual evolution of a sum of ideas – not only Leibniz’s, but also Boscovich’s, Kant’s, Faraday’s and Maxwell’s (See Gale 1974, 340) – within the content of this article I will only pursue the Leibniz-Chew axis of conceptual symmetry, illustrating how Leibniz’s metaphysics *mirrors*

Chew's physics and vice versa, a fact that allows us to consider both of them self-consistent.

There are three striking similarities through which Chew's physics and Leibniz's metaphysics mirror each other: (a) the anti-corpuscularian entities based on the energetic system of force centres, circumscribed to an ontology of forces, (b) the relations of symmetry between hadrons and "mirroring" between monads, based on the relations of intercorrelation or mutual causality, "whereby if x is a cause of y , then y is also a cause of x " (Gale 1974, 340), (if $x \rightarrow y$, then $y \rightarrow x$), (c) the logical criterion of mutual, reciprocal self-consistency, according to which "given one, the whole system is given and vice versa" (Gale 1974, 348).

(a) *The first similarity*

Let's start by noticing a very interesting thing about the anti-corpuscularian ontology common to Leibniz and Chew, who are part of the same family in the dispute between substantialists and antishubstantialists (the famous dispute between Leibniz and Newton later becoming the contemporary dispute between corpuscularian and anti-corpuscularian ontology). George Gale argues in this regard that "the new anti-corpuscularian physical ontology is not as radically heterodox as it appears. In fact, the contemporary alternative to corpuscularianism, at least as exemplified by Geoffrey Chew, one of its leading proponents, is interestingly analogous to the anti-corpuscularian ontology suggested originally by Leibniz" (Gale 1974, 339).

The anti-corpuscularian ontology denies, as we have seen, the existence of fundamental constituents, namely corpuscular entities, based on an energy system of central forces or interactions between forces – that is, based on the concept of force circumscribed to an *ontology of forces* common to Leibniz's Monadology and Chew's Bootstrap theory. Leibniz calls this primitive or anti-corpuscularian entity based on centres of force or energy, *monad*, while Chew refers to it as *hadron*.

We know that for Leibniz, in contrast to Descartes, extension is infinitely divisible; hence, the concept of a unity (i.e. the unity of an object) is incomprehensible, and thus the concept of material substance is as well. Given this, the immaterial substance of things becomes for him the energy, the force, that internal potency, *vis viva*, with which

every monad is endowed. Only the monad, the living unit, the non-extended/non-quantitative energy atom, simple and without parts, i.e., indivisible, can enter into the composition of compounds – *aggregatum* (See Leibniz 1898, §§ 1-3, 217-8). Analogously, the primitive, anti-corpuscularian entity from Chew's physics ontology is a "strong force" called hadron. Hadrons describe strong interactions between particles which form the nucleuses of atoms, namely neutrons and protons. According to Chew's bootstrap vision, hadrons (neutron and proton) in themselves are not understood as fundamental, being nothing more than two particular states of matter which interact strongly. Thus, states Chew, "all these particles are dynamical structures in the sense that they represent a delicate balance of forces; indeed, they owe their existence to the same forces through which they mutually interact" (Chew 1970, 23). These hadronic interactions are completely different in comparison to traditional interactions, i.e., gravity, weak force, etc. The reactions take place through scattering collisions, in which two hadrons collide and go off in different directions. If a hadron is moving with enough energy before striking a hadron at rest, a new hadron can be created in the collision. "This is a significant point of analogy between force conceptions of both Leibniz and Chew: *immaterial forces are the source of materiality*. For Leibniz, the active and passive primitive forces in corporeal substances ground, and in some sense, produce the observed properties of matter. This is quite similar to the production of a third particle by means of strong-force interactions between two other parties" (Gale 1974, 344). Thus if the basic forces are interactions that can *create*, annihilate and transform particles (See Chew 1964, 79) then, for both Chew and Leibniz, forces produce, or create particles on the basis of equivalent implication or reciprocal causality relations, which Chew has also called "crossed reaction", according to which, *if forces are interactions that create particles, then the particles are forces*. Given that, according to the bootstrap hypothesis, the strongly interacting particles are all dynamical, energetic structures that owe their existence to the same forces through which they mutually interact, "the bootstrap system may be viewed, therefore, as a physical interpreta-

tion of the dynamical aspects of Leibniz's Monadology" (Gale 1974, 345).

(b) The second similarity

The second similarity between Leibniz and Chew is derived from an analysis of the first. Since forces are (immaterial) anti-corpuscularian entities, this makes possible the relations of symmetry between hadrons, or the mirroring between monads, on the basis of reciprocal implication causality (if $x \rightarrow y$, then $y \rightarrow x$), which Chew has called crossing symmetry. This also seems to be the implication when Chew, in reference to equivalent readings of a crossed reaction diagram, states that "a single diagram corresponds both to a force in one reaction and to an intermediate particle in the crossed reaction" (Chew 1964, 92), indicating that the particle bears a symmetrical relation to or mirrors the force and vice versa. Thus, when one reads the diagram, convenience and *perspective* alone determine whether one calls the entity involved a "particle" or a "force" (See Gale 1974, 343).

The study of subatomic particles and strong interactions has highlighted, therefore, that the universe is not composed of fundamental properties as long as particles are simultaneously forces and vice versa, and as long as the differences between them – meaning their unicity – derive just from the position of perception or observation (as is also true of Leibniz's monads).

These strong interactions of subatomic particles, far from being chaotic, reveal a *pattern* – a relation of symmetry that gives rise to observable *regularities*. If "in atomic physics they can be explained in terms of the properties and mutual interactions of the atom's constituents (the protons, neutrons and electrons), in particle physics such an explanation has not yet been possible" (Capra 1975, 248-9), given the fact that here, objects composed of elementary parts no longer exist. Experiments involving particle accelerators have demonstrated that when splitting particles into elementary parts, the result is not "smaller pieces of the original particles", indicating that the structure of a subatomic particle cannot be understood in the static sense of it being composed of elementary parts, but should rather be understood in the dynamic sense of properties and interactions (Capra 1975, 249). The process of "breaking particles up" results in

particles of the same kind as the original ones; from this it follows that at this level only regularities can be described – regularities grouped into "families" in which the notion of *symmetry* plays a decisive role.

If a notion of symmetry is generally applied to the reflection in the mirror, in subatomic physics it is applied *simultaneously to particles and interactions*, to which they are inseparably tied. As any interaction which presents a symmetry implies invariance, symmetry at the level of particle properties translates through the four laws of conservation⁶ within the interaction processes (See Capra 1975, 252). Symmetry and conservation describe, therefore, *regularities* observed at the subatomic particle level. Even if hadrons present quark symmetries, behaving as if they were composed of elementary constituents, "self-standing quarks" have yet to be found, whereas symmetries and regularities abound.

Subsequently, the hadron, analogous to the monad, is constituted by other hadrons, which it in turn constitutes. This network is possibly infinite in its extent. Moreover, on this view, "each single hadron can be, and probably is, individual, with its individuation being given by its precise *mirroring of the total situation of all other hadrons, particularly those which immediately determine it*" (Gale 1974, 345). Each of these latter particles (hadrons) in turn owes its existence to a *set of forces* to which the first particle makes a contribution. In other words, each particle helps to generate other particles, which in turn generate it. In this circular or non-linear situation, it is possible to imagine, says Chew, "that no free, or arbitrary, variables appear and that the only self-consistent set of particles is the one found in nature" (Chew 1964, 93).

(c) The third similarity

Self-consistency, the core of Chew's bootstrap theory and Leibniz's Monadology, represents the third striking similarity, as it is the logical criterion of the two systems whose circular construction comes to prove that *the universe is One*. This self-consistency, based on the mutual intercorrelation of

⁶ We are referring to 1) symmetry in relation to the position in space, 2) in time, 3) in relation to the orientation in space and 4) in relation to the electrical charge (See Capra 1975, 252).

things highlights, therefore, a unitary and indivisible universe in which all objects, phenomena and events are intimately correlated. This intercorrelation is not imposed by a force, will or external law⁷, but derives from the *internal basis* of the fact that all are parts of an indivisible whole whose *internal law* creates *self-consistency* (with itself and all other parts possible) on the basis of the mutual part-whole equivalence.

By highlighting this universal interconnection, the bootstrap conception, like Leibniz's Monadology, demonstrates that explaining a phenomenon refers to explaining its connections with the whole, since the properties of the parts are not determined by a fundamental law, but by the properties of the other parts of the whole (See Capra 1975, 290), which constitutes self-consistency. Given that the properties of the part are not static, but dynamic, which is to say the consequence of its interactions with the other parts and with the whole, we can then state that, not only in Leibniz's metaphysical sense, but in the physical or scientific sense of the bootstrap theory, the *part* is in fact *the whole*, "containing" and mutually implicating the whole – and vice versa.

This relation of "mirroring", symmetry or part-whole implication was validated by the *bootstrap* conception regarding the behaviour of hadrons. Thus, hadrons "contain" each other in the sense that they mutually "implicate" each other, since hadrons are structures also made up of hadrons, in which none is more "fundamental" than the others. As the forces which maintain these structures manifest through the exchange of particles in which the particles that make the transfer are also hadrons, this means that the whole set of hadrons self-generates and continues indefinitely through its connections (See Capra 1975, 296), thus resulting

in a self-determined and self-consistent mechanism in which all the particles are composed of the others, or mutually "contain themselves", such that a particle contains the others and at the same time is contained by them. This means, according to Chew, that the Universe is *all of one piece*, which is to say *One*, in the sense that *for any hadrons to exist at all, the entire family must exist in a mutually supporting framework*. From this perspective, as Gale pointed out, the hadron universe is not evolutionary, or linear, but circular – it must come into existence *all at once*, since the individuals are given only in the mirroring relations they have with each other. "Since each hadron is composed of others, which in turn it composes, they must all be given at once in the dynamical system. We cannot imagine our particular hadronic universe as being different from what it is, since, if even a single particle were different, all the other particles must indeed be different, due to the reciprocal mirroring quality of their compositional relations" (Gale 1974, 346).

This physical conception is similar to Leibniz's metaphysical idea of *monads*, each of which "mirrors" the whole Universe in its own way, some in great detail and others rather vaguely. "Thus, although each created monad represents the whole universe, it more distinctly represents the body which especially pertains to it, and of which it is the entelechy; and as this body expresses the whole universe through the connection of all matter in the *plenum*, the soul also represents the whole universe in representing this body, which belongs to it in a special way" (Leibniz 1898, § 62, 253). Leibniz thus intuitively, at a metaphysical level, that in the smallest portion of matter the whole, the world, the universe is given: "Each portion of matter may be conceived as like a garden full of plants and like a pond full of fish. But each branch of every plant, each member of every animal, each drop of its liquid parts is also some such garden or pond" (Leibniz 1898, § 67, 256). These "portions of matter", of which Leibniz speaks here, are ultimately monads, and every monad *must ideally contain the whole universe*. The monads are infinite in number, and each, as it ideally contains all, *must* therefore contain an infinity of "parts".

Thus, analogous to the system of existing hadrons, the system of existing monads is also the set which is maximally self-consistent *as long as given*

⁷ Joseph Needham shows in this sense that the strictly Western notion of a fundamental law of nature derives from the (monotheistic) Judaeo-Christian faith in a divine legislator. It is this very belief, so rooted in the Western mentality, which made it possible, three centuries after Newton, for the main goal of the natural sciences to consist in searching for and postulating the ultimate laws of nature. Regarding Chinese cultural influences on Leibniz, Needham maintains that Leibniz had familiarised himself with Chinese spirituality through translations made by Jesuit monks, and that it is quite possible that his philosophical system was inspired by the neo-Confucian school (See J. Needham 1956, vol II, 496).

any monad, all other consistent monads are given and must be given (See Gale 1974, 346). “It may be said that a monad can only come into being or come to an end all at once; that is, it can come into being only by creation and end by annihilation” (Leibniz 1898, § 6, 219). This means that, for both Leibniz and Chew the “laws of nature” are generated in the same way, being necessarily as they are, since they are determined by the behaviour of the set of entities – either hadrons or monads. The difference between Chew and Leibniz, then, is that if for Leibniz “this world is the best of *all possible worlds*”, for Chew this universe is unique, being *the only possible* universe – excluding any other possible worlds for lawfulness.

Beyond this difference, it is clear that the criterion of self-consistency, common to both systems, is so striking that, as Gale states, “the reciprocal compositional relation which generates the hadrons is a physical interpretation of the intermonadic relations” (Gale 1974, 347). Thus, Chew’s system made explicit a feature of Leibniz’s monadic system which was quite probably “at the bottom of Leibniz’s inability to discover a physical dynamic that completely interpreted his metaphysical system” (Gale 1974, 347).

In order to understand these intermonadic relations of mirroring, which describe the self-consistency of the monadic system, one must firstly understand why “the monads have no windows, through which anything could come in or go out” (Leibniz 1898, § 7, 219). Why indeed are Leibniz’s monads mirrors without windows, which do not communicate on the outside, but only on the inside by virtue of an internal principle? Firstly, it seems quite clear that if this basis were not internal, but were imposed from the outside through a force, law or external will, the monadic system would cease to be self-consistent with itself and with all the other monads within the system. Secondly, the monads do not have windows because nothing can penetrate and “add” something, except perhaps that *addition of nonbeing* (which Plotinus mentions), the monad already being a *complete idea*⁸, which is to say self-consistent with itself and with the monadic system precisely by virtue of the fact that the monad is a

perception of the whole universe, or “a perpetual living mirror of the universe” (Leibniz 1898, § 56, 248). As a complete idea, the monad contains or mirrors in its interiority, i.e. in its present or actuality, all the past and future/potential states of the universe. Through this internal mirroring, the intermonadic “communication” is limitless, because it can extend up to any distance in time and space. “It follows that this inter-communication of things extends to any distance, however great. And consequently every body feels the effect of all that takes place in the universe, so that he who sees all might read in each individual one what is happening everywhere, and even what has happened or shall happen, observing in the present that which is far off in time as well as in space” (Leibniz 1898, § 61, 251-2).

Therefore, *monads do not have “windows” to the exterior precisely because they have “mirrors” of the interior* through which they communicate spontaneously, mutually, globally and circularly with the entire universe, each monad mutually implicating the others in the sense of reciprocal and circular self-consistency, according to which *if a monad is given, all the other ones – which is to say the entire system – are given, or are mirrored all at once*. In other words, each monad has an instantaneous connection with the whole universe/monadic system, which it reflects more or less clearly from its point of view, this universe being *One*, indivisible and self-consistent as long as its laws necessarily spring from within by virtue of the mutual part-whole intercorrelation.

4. Instead of a conclusion: Leibniz's boots and Chew's straps or Bootstrap's Monadology

It is interesting to notice that according to the Cambridge Dictionary, *bootstrap* is a small and strong strap of a boot that enables you to pull the entire boot on. Beyond the concrete meaning of the word, there is also a metaphorical one, extremely suggestive as it appears in expressions such as: “pulling yourself up by your own bootstraps” or “pull oneself up by one's bootstrap”, meaning to help oneself without the aid of others using *one's forces or resources*. Thus, the term bootstrap refers from the very beginning to the idea of one's *own*

⁸ Like Plato, Leibniz believed that all our ideas are already within us; we “learn” nothing, merely “reminisce”.

force or internal basis, that is to the idea of self-sufficiency, self-containment and self-determination, which implies self-consistency, the hard central core of Chew's *physical* bootstrap, as well as Leibniz's *metaphysical* one.

Taking into account, the fact that “we cannot expect people to lift themselves up by their bootstraps when they do not even have boots”, we can understand that without Leibniz's metaphysical “boots” i.e., without this epistemological heritage, Chew's “straps”, his own forces, no matter how strong, would not have been able to propel his vision of the bootstrap perception to the height of the present day. Similarly, Leibniz's “boots” have today become so significant, modern and up-to-date because of the force of Chew's “strap” propulsion, which enriched, clarified and established at a physical level the vision of Leibniz's Monadology, so that, by virtue of the mirroring/symmetry connections, the two levels, the physical and metaphysical, can be considered self-consistent, thus generating a new model – the *Bootstrap's Monadology model*.

In other words, due to the intercorrelation of the two levels of reality, the physical – matter – and the metaphysical – consciousness – we can say without exaggeration that Chew's physics and Leibniz's metaphysics are intertwined, interconnected and mirrored, like twin souls, so much so that it can be stated that if Chew's bootstrap theory represents the *physical* explanatory level of Leibniz's metaphysics then, based on the same mirroring or symmetry connections, Leibniz's Monadology represents the metaphysical explanatory level of Chew's physics.

Given that Leibniz's metaphysics is mirrored symmetrically in Chew's physics and vice versa, in the form of a crossing symmetry, the bootstrap's monadology model can be considered coherent and self-consistent as long as it reflects the unity or self-consistency of the physical level of *matter* with the metaphysical level of *consciousness*, validating, therefore, an ancient Hermetic law according to which: “As above, so below; as below, so above” (Carty 2007, 65). This Hermetic law is probably the most eloquent, elegant, simple and fascinating example of mirroring, symmetry, correspondence, self-consistency and unity.

References

- Ardelean R (2020) The Theoretical Legitimation of an Ethical Paradigm of Human Consciousness Nonseparability starting from the Scientific Paradigm of Quantum Theoretical Nonseparability. *Hermeneia. Journal of Hermeneutics, Art Theory and Criticism* 25: 60–67, Iași: Axis Academic Foundation.
- Bohm D (2002) *Wholeness and the Implicate Order*, London and New York: Routledge Classics.
- Bohm D, Hiley B J (1975) On the Intuitive Understanding of Non-Locality as Implied by Quantum Theory, *Foundations of Physics* 5: 93–109. https://www.researchgate.net/publication/226296304_On_the_Intuitive_Understanding_of_Non-Locality_as_Implied_by_Quantum_Theory.
- Bohr N (1934) *Atomic Physics and the Description of Nature*, London : Cambridge University Press.
- Capra F (1975) *The Tao of Physics. An Exploration of the Parallels Between Modern Physics and Eastern Mysticism*. Boulder, Colorado: Shambhala Publications.
- Carty G D (2007) *The Emerald Tablet: And the Alchemy of Spiritual Transformation*. Atlanta, Georgia: Personal Development Institute.
- Chew F G (1968) Bootstrap: A Scientific Idea? *Science* 161(8): 762–765, US: American Association for the Advancement of Science (AAAS).
- Chew, F. Geoffrey.1970. “Hadron Bootstrap: Triumph or Frustration?”, *Physics Today*, Vol. 23, No.10, pp.23-28, American Institute of Physics: AIP Publishing.
- Chew F G, Gell-Mann M, Rosenfeld A H (1964) Strongly Interacting Particles. *Scientific American* 210(2): 74–93, US: Scientific American, a division of Nature.
- Gale G (1974) Chew's Monadology. *Journal of History of Ideas* 35(2 Apr.-Jun.): 339–348, US: University of Pennsylvania Press.
- Heisenberg W (1963) *Physics and Philosophy*. London: Allen & Unwin
- Leibniz W G (1898) *The Monadology and Other Philosophical Writings*, translated with Introduction and Notes by Robert Latta, London: Oxford University Press

Needham J (1956) *Science and Civilisation in China*, Vol. II, London: Cambridge University Press.

Shimony A (1988) The Reality of Quantum World. *Scientific American* 258(1): 46–53, US: Scientific American, a division of Nature.

Book reviews

Gunnar Skirbekk. *Religion in Modern Society (Religion i moderne samfunn)*.

Dreyers Forlag Oslo. 2021

[Reviewed by Silviya Serafimova, Institute of Philosophy and Sociology – BAS,

email: silviya_serafimova@yahoo.com

Gunnar Skirbekk's thought-provoking monograph, *Religion in Modern Society*, raises the burning question of how we should think about religion in a "science-based and institutionally differentiated society in crisis" (p. 173). The necessity of philosophically clarifying what one understands by religion, as refracted through the lens of issues such as religion-based social integration, points toward the search for an ontologically grounded explanatory framework. In turn, the latter is underscored by another set of fundamental questions, such as that about the status of religion in modern society as a matter of adopting *either-or* or *both-and* modes. This means that one can tackle the conceptual assumptions of arguing for religion *either* as a problem *or* as a resource, or rather as a matter of being *both* a problem *and* a resource (Ibid.).

For the purposes of clarifying the ontological pressure behind the complexity of the aforementioned modes, Skirbekk examines the epistemological, institutional claims and the role of crisis in regard to what he coins adjustment (*tilpassing*) (p. 174) — a necessitated adaptation of both religious and non-religious people, as well as believers in different teachings, to the fundamental conditions of the modern, science-based and institutionally differentiated society in crisis. It is the precarious complexity of the society in question that renders the urge to analyze the genealogy of its epistemic, cultural and institutional diversity a key for understanding the conceptual complexity of the analysis of religion itself.

In this context, Skirbekk cogently points out the role of the epistemological claim of having no phobia of arguments ("vi må ikkje bli "argumentofobe") as complementing that of not being "partially modern" ("vi må ikkje bli "halvmoderne") (Ibid.). In a sense, one may argue that the aforementioned two claims provide a diagnosis of how a society in crisis can overcome it by fostering global solidarity as a remedy for the conflicts between cultures. In other words, epistemological claims can be interpreted as justifying the role of religion *as a resource, not as a problem*, viz. one may predict the overcoming of society's crisis by using religion as a *resource* for social integration, thus avoiding the *problems related to* intercultural struggles.

Regarding institutional claims, Skirbekk sees their contribution mainly in the following respects — establishing functioning modern institutions as a matter of fulfilling one political and socially meaningful objective and behaving such that different institutions can become foreseeable that fulfill personal goals. In turn, the challenges of adjusting different public and personal roles, which are supposed to be kept separate, lead to what Skirbekk calls knowledge-related and institutional challenges. The latter apply pressure upon both religious and non-religious people. Consequently, such challenges become existentially exclusive in times of crisis, which necessitates the recognition of religion as a resource. What is of crucial importance in this context is the way in which Skirbekk outlines the ethical framework of dealing with the

aforementioned challenges. He wisely argues that they should not be evaluated through the lens of situational ethics (“sinnelagsetikk”), but rather through that of consequentialist ethics (“konsekvensetikk”) (p. 175).

The monograph *Religion in Modern Society* consists of four main parts which provide multi-disciplinary insights from the perspective of the theory of science in philosophical terms (i.e. vitkapsfilosofisk perspektiv) (p. 11). While the first part is devoted to the clarification of the relationships between science and religion refracted through the lens of the theory of science in modern science-based societies, the second one is focused upon the relationship between religion and freedom of speech in modern states ruled by justice. In turn, the objective of the third section is to shed light upon religious policy in modern welfare states, while the fourth part displays a comparative analysis of some theoretical contributions, including those of Habermas, Ahmet T. Kuru and Skirbekk himself.

Specifically, the first part focuses on clarifying eight points related to the role of science and religion in the modern science-based and crisis-framed society. These include a principle which is commonly shared by all sciences, viz. the lower value of instrumental use at the expense of one “informed and self-critical” (p. 19) set of arguments implied as a matter of mutual learning processes (point one). The second point explicitly addresses one of the aforementioned epistemological claims — that of avoiding the status of being “partially modern”. The reasons behind this point are related to the necessity of having an informed and self-critical approach to different subjects which are tackled by different sciences in modern society. Maybe one of the most important points in the debate is the third point, which sheds light upon the issue that scientific pluralism is not axiologically neutral. That is why it requires the pluralism in question to be rooted in an “internal ethos of improvement” (p. 22). The improvement itself is understood as addressing what the potentially valuable in knowledge is by avoiding the “less reasonable” and encouraging what is “more reasonable” (p. 24). The axiological implications of the third point clearly make room for the normative justification of the fourth point, according to which science is cogently defined as being both part of the problem and part of the solution. This means that although science cannot by itself solve the complex and burning dilemmas of the modern society in crisis, it is the commonly shared self-critical

potential of the sciences that can contribute to revealing the genealogy of such problems.

In this context, Skirbekk contextualizes the role of religion, as displayed by the last three points. Understanding the diversity of religions is impossible unless one takes into account the plurality of the sciences and that of different cultures. Similarly to cultural pluralism and scientific pluralism, religious pluralism also requires some conceptual specifications to be provided. According to the fifth point, the pluralism of religions necessitates the introduction of clarifying definitions, since one cannot argue for respect and support merely because something has been called a religion. By conducting a comparative analysis between science and religion in western history, Skirbekk formulates the sixth point — that modernity sets the grounds of a positive relationship between religion and science, not least among the natural scientists. In turn, the seventh point addresses “the internal necessity of a critique of religion” (p. 32), which is typical of the three main monotheistic religions. The last point as such can be addressed to the third one regarding the adjustment provided by the internal ethos. This time, however, the analysis is conducted in terms of the way in which a clarified understanding of religion can play a positive role if ‘adjusted’ to the modern society in crisis. Elaborating upon Habermas’ concept of the modernization of consciousness, Skirbekk provides a genealogical analysis that demonstrates why the three main monotheistic religions are confronted with the same fundamental challenges today regarding institutional adjustment and adjustment of knowledge in order to coexist with positive and necessary claims that characterize an enlightened modernity (p. 36).

Examining the complexity of the reasons behind the process of modernization is of crucial importance for understanding how the discursive analysis of the problem of evil is related to the genealogy of the narrative about God. The multiple embodiments of this narrative are outlined by an examination of the following five points. First, the relationship between faith in God and the vision of a religious foundation for morality is displaced in time, which requires one to track the changes in the socio-cultural and political conditions leading to this displacement. The second point addresses such changes in process terms by outlining the differences in the way that the three main monotheistic religions recognize the concept of God. All the

aforementioned specifications necessitate the introduction of the third point regarding the narrative about God, viz. the problem of evil, as refracted through the lens of the three religions in question. The last two points, point four (“What does it mean to talk about God?”, as a tyrant, as being stupid, or as being indifferent to the choice of our deeds) and point five (“What does it mean to talk about the Holy?”, as affecting the vulnerable and perishable) are raised as questions which make room for a discursive analysis of the critique of religion as such.

On a macro-methodological level, the discursive analysis first addresses the epistemological claims, viz. claims about the importance of avoiding the behavior of being “phobic of arguments”, as well as being “partially modern”. These claims are raised in response to the question “Why do we need a critique of religion?” Consequently, the question of “How should we do this?” addresses the symbolic capital of practice as a cornerstone of this critique. In order to achieve it, Skirbekk points out the role of education and textbooks. In this context, the critique of religion should be examined as a process of clarifying, as a way of modernizing consciousness, as well as pointing towards further perspectives of investigation.

The second part is devoted to the exploration of practical challenges that derive from the normative justification of religion within the modern society in crisis. The analysis of these challenges illustrates why the conceptual clarification of what one understands by religion (due to the specific socio-cultural and political discourse where the critique of religion takes place) illuminates how religion can turn into a stimulus of social integration or social disintegration. By examining the Jyllands-Posten Muhammad cartoons controversy (p. 64), Skirbekk examines the issue of freedom of speech through the lens of three “insufficient basic strategies” (p. 63). The first one, related to the usefulness of having a legal principle of freedom of speech, raises the question of what one should do when it is not useful to operate with such a principle. The second strategy addresses concerns about the legal delimitation of the fundamental principle of individual freedom, as well as how it is considered a Western value in the modern globalized world. This, however, does not make it a universally valid principle. Skirbekk also cogently points out that the third strategy regarding decisionism cannot be recognized as a grounding principle because it provides the introduction of apparent methodological relativism. Furthermore, he

discusses in detail why freedom of speech is not “merely one Western value” (p. 65) nor is it one among many other values. Skirbekk points out that not every single statement should be uncritically protected by the freedom of speech. In addition, he also provides a ‘positive’ definition of what freedom of speech should look like, viz. as a constitutive assumption of a discursive searching for truth.

The third part devoted to the clarification of the role of religion for modern institutions is mainly focused upon the different embodiments of multiculturalism in the welfare state. Similarly to the way in which Skirbekk reveals the multidimensionality of the issue of religion, he provides a detailed conceptual revision of multiculturalism. Skirbekk outlines four main specifications. The first and the second ones concern the boundaries of multiculturalism in cases in which there is a close relationship between social institutions and cultural identity, as well as cultural diversity in a science-based society. The last two points address the religious defense of multiculturalism from a science-based perspective and consider potential conflicts that can arise when evaluating multiculturalism and the welfare state in long terms.

Analyzing the hypothesis that the welfare state in Scandinavia has roots in the headstrong interplay between national variants of Lutheran Protestantism and the successful people’s movements of the early 19th century, Skirbekk examines historical perspectives and those of the theory of modernity as they relate to cultural change. The institutional development and learning processes of the 18th century are considered as charting the progress of state institutions and the based upon knowledge political initiatives, with a focus upon the power of law, the ideal of Enlightenment and the role of the people, self-organization and socio-political learning. Specifically, the Danish-Norwegian variant of Lutheran Protestantism played a crucial role during the 18th and 19th centuries — first, as a state institution dealing with the external processes of enlightenment and discipline, and second, as an internal inspiration for popular people’s movements (folkelege rørsler) channeling people’s enlightenment (folkeopplysningen) and self-organization (p. 112).

In this context, Skirbekk outlines the complex dynamics triggered by the interplay between the educated at the university state officials and the engaged and enlightened laypeople. The latter, in

particular, drive the socio-cultural processes of modernization which make room for parliamentarism. Skirbekk cogently points out that multiculturalism in a modern vulnerable (sårbare) society (p. 114) — for instance, in the Nordic societies — makes room for a “self-critical critique of religion” (sjølvkritisk religionskritikk) (Ibid.). Consequently, the critique in question contributes to processes of examining the complexity of multiculturalism as a part of contemporary Nordic democracies by revising the role of religion as a key for social integration.

However, Skirbekk also explains how numerous challenges in the Nordic welfare state stem from two interrelated systems, viz. that of the liberal welfare state and that of global, technologically based capitalism. The issue of multiculturalism is tackled from the perspective of the theory of science, which raises four main points. The first concerns the close relationship between social institutions and cultural identity. The second addresses the issue that being a modern, science-based society, the Nordic welfare state should display a variety of disciplinary approaches and institutional differentiations which make cultural diversity “a special form of multiculturalism” (p. 106). Furthermore, this modern society should support the arguments of gaining clarity about the value of religion by engaging in reasoned discussions before passing judgement. This third point is of crucial importance, since it opens critical discussions about how the concept of God is interpreted and thus, how these interpretations may be reformulated and modernized. In turn, the fourth point concerns evaluations related to the variable of time, viz. the long-term and short-term implications of multiculturalism in the Nordic welfare state should be examined with the goal of assessing whether the future will be peaceful and harmonious or instead subject to serious crises and conflicts.

What are some of the most important implications of the aforementioned four points? Skirbekk cogently outlines the role of the common ethos which should guide reflections and critical discussions about multicultural diversity. In other words, the cultivation of such an ethos in the modern Nordic welfare state, being based on science, as well as being democratic in nature, can contribute to the enlightening project of discussing the role of multiculturalism “in a reflexive and discursive manner” (p. 109). Furthermore, elaborating upon Habermas’ three-step argument for the modernization of consciousness, which is informed

by (1) the recognition of religious diversity, (2) the recognition of scientific insights and forms of work that are consistent with the scientific critical and (3) the recognition of the institutional difference between religion and juridical system/politics, Skirbekk clarifies why religion cannot be separated from the roles of science and reason in modern society.

In the fourth and last part of his monograph, Skirbekk comparatively examines the theoretical reasons behind two different types of religion’s embodiments — that of the historical contextualization of religion in secular society, which represents the gradual modification of the relationships between knowledge and faith (as displayed by Habermas’ interpretation), and that of religion refracted through the lens of class politics, the class perspective which reconsiders the relationships between essentialism and post-colonialism (as represented by Ahmet T. Kuru’s theory). Skirbekk examines Habermas’ objective of investigating religion as a meaningful world picture which is evaluated in a socially integrating perspective. Elaborating upon Habermas’ concluding remarks, Skirbekk discusses the role of the learning processes which address the world picture’s structural changes. They are refracted through the lens of a double-bind approach, viz. as encouraging hope for “reasonable freedom” (håp om “fornuftig fridom”) (p. 153), as well as facilitating a degree of openness for a “translation” (omsetjing) (Ibid.) of religious experiences into a secular world.

Skirbekk himself does not uncritically accept Habermas’ theory, but uses it as a springboard to analyze a variety of issues that deserve further investigation, such as neglected climate-related, institutionally-related and knowledge-related challenges facing contemporary societies. Revising these issues is essential for building relevant time diagnosis which should grasp the complexity of reality. This points to a crucial concern that Skirbekk identifies, viz., that pre-modern and politicized versions of religion are not always favorable to solidarity and may in fact drive wedges intergenerationally or internationally among members of the same generation.

This clarification is of primary importance because it makes room for reevaluating the influence of religion in its current social role. Specifically, such a reevaluation is achieved when we reconsider the dialectical relationship between the current and future impacts of religion. Broadening the scope of

one's investigation to include the time diagnosis in this way may allow us to more accurately predict which forms of religion are potential triggers for conflicts and social disintegration. In this context, Skirbekk wisely reiterates the necessity of providing more precise definitions of religion since without precision it is impossible to counteract the trend of decreasing social integration in today's complicated globalized society. The practical embodiments of such a reminder can be seen in the way in which one can discern between "similar and dissimilar, between religions (or religious variants) which are characterized by a "modernization of the religious consciousness" and religions which are not" (p. 159).

In turn, Skirbekk's analysis of Ahmet T. Kuru's interpretation of Islam, as examined from a historical class perspective, points toward a critical revision of the way in which Islam and progress stop 'going hand in hand'. Kuru's interpretation, as displayed by Skirbekk, argues that this interruption in progress in the Muslim world is caused by a repressive class constellation, by an alliance of religious and military leadership, and not by Islam per se; hence, this kind of 'essentialist' critique of Islam is rejected. Similarly, Kuru argues that Western colonialism came later than the first decline in the Muslim world; hence, this kind of 'post-colonialist' explanation is also rejected.

What Skirbekk considers to be questionable in Kuru's interpretation is the relatively scarce analysis of modern society and the world today, as well as the discrepancy between Kuru's stated purpose of critical interpretation and the book's emphasis on a historical review. Comparing and contrasting Kuru's and Habermas' theories, Skirbekk reaches the conclusion that Kuru's book can be read in parallel with Habermas' work, thus complementing Habermas' understanding of social integration with an analysis of social disintegration.

In his thought-provoking monograph, Skirbekk

interprets the problem of religion in modern society as *modus vivendi* since, like life itself, religion is at once remarkable and vulnerable. Similarly to the way in which Aristotle recognizes the beginning of philosophy in the process of wondering, one may say that Skirbekk sees the beginning of religion in the act of wondering in both existential and cosmological terms; the soul of religion lies □ not in any particular answer given here and now, but in the process of seeking humanity's essence, which is conceived of as inseparable from that of the cosmos. That is why the process of wondering, which is irreducible to a given set of questions and answers, is naturally anticipated by Skirbekk as a process of reflection upon the biggest questions. Such questions include the problem of evil and its implications for human nature. This is due to the fact that initially, the language of religion is not a dogmatic language, but a language expressing both sorrow and wonder — a language, which according to Skirbekk, shows "respect and reverence for vulnerable life" (p. 177). Furthermore, the topic of religion as *modus vivendi*, which addresses the reverence for life, is expressed through thoughtful engagement, which encourages the process of critical reflection.

Sharing the complex process of philosophical reflection with his readers, Skirbekk shows the development of his own critical engagement with the topic, which dates back to Skirbekk's first attempts at 'wondering' as a young philosopher, as displayed in his book *Nihilism? (Nihilisme?)* (1958)), matures as Skirbekk becomes established as a world-recognized philosopher with expertise in philosophical 'wondering' (*Wonders (Undringa)* (2020)), and culminates in the monograph *Religion in Modern Society*. This inspiring monograph demonstrates why it is in times of crisis that the philosophical *modus vivendi* is one of the few ways that can bring us back to our human nature, guiding us to be both sensitive and responsible for the life we (*should*) live together with other beings.

BOOK REVIEW of Röck, Tina (2022) *Dynamic Realism Uncovering the Reality of Becoming through Phenomenology and Process Philosophy*, Edinburgh University Press

Marina Bakalova

Institute of Philosophy and Sociology, Bulgarian Academy of Sciences

e-mail: marina.bakalova@gmail.com

The book *Dynamic Realism Uncovering the Reality of Becoming through Phenomenology and Process Philosophy* is a return to the grand questions of philosophy and the hallmark speculative methods of the 20th century: Husserl's phenomenology and Whitehead's process metaphysics. It is an attempt to combine both methodologies in a unified approach. The author, Tina Röck, says that the book is an attempt to come to terms with the contemporary world, in which the immense speed and complexity of information technologies threatens to overtake physical reality (as we conceive of it). The motivation comes from the author's conviction that philosophy should take into account the situational context of real everyday life. Thus the main question she raises is this:

If this everyday world that we live in is to be taken to be what is real (as our actual, direct experience would suggest), what would be the hallmark properties that characterise all of what there is? In other words, given these premises, how should we characterise what there is? (p. 61).

The first general observation of Tina Röck is that our experience is overwhelmed with dynamism and change. In her words, there is "no such thing in inner or outer experience as timelessness" (p. 29). Hence, in her view we cannot adequately explain our current situation by means of reason and logic, or any sort of analytic method that aims at spelling out things in a stable and hence atemporal way. In contrast to the established methods of research, she argues, we need an approach that is adequate to the dynamic changes in the world, and which integrates not just thought, but experience with its typical qualities. This would enable us to explain human interaction with the complex and

unstable environment. Tina Röck sees the marriage of phenomenology and process ontology as the best way to do so:

Basing metaphysical speculation on real possibilities grounded in and guaranteed by (phenomenological) descriptions of the spatiotemporal world as given in experience thus allows for a form of speculation that does not immediately fall prey to the lure of intellectualism or rationalism, since it is not bound by the possibilities of thought alone (p. 43).

This short quote shapes the context of the study, and raises the following question about its premises. Why should one base a philosophical investigation on contingent characteristics of how reality appears to us, and make ontological commitments based on that? This, I suppose, would be a worry for many philosophers, and without further argument they would not see sufficient reason to reject the traditional methods of philosophical investigation. An answer to such a worry is not explicitly given in the book. However, one may guess that it is rooted in the deep convictions of process philosophers that there is a dynamic and creative element in our primary experience of the world which is of such great value and importance to us that it could and should ground an alternative approach to metaphysics.

Next, given the author's disapproval of rationalistic methodology, there is an issue here of how to avoid the method being as messy as the reality is in order to grasp what is creative and valuable for us. Tina Röck answers that we have to find a balanced approach – a way paved with certain rules and order and yet open to the dynamic and the creative elements of our experience that potentially also

reflect something in reality. The book is an attempt to open such a middle ground by clarifying specific categories related to temporality, change and dynamicity. Among the categories discussed in the book is the category of adequacy, which determines new standards of acceptable certainty and allows for degrees of adequacy (pp. 34-37); the category of 'becoming existence' which substitutes that of "being" (pp. 203-208), etc. The exposition covers different authors and conceptions related to the project. Special attention is given to the philosophy of Martin Heidegger who prepared the path for a fruitful combination of ontological and phenomenological investigations in his *Being and Time*.

The book is all about correlation between thinking and being, and the marriage between phenomenology and process philosophy is meant to clarify exactly this correlation:

"In the phenomenological tradition experience in general is understood as an intentional act that allows the inner as well as the outer world to be there for us" (p. 71).

The author's main claim is that reality is fundamentally temporal and thus changing or processual, because the aspects of experience that show the world are changeable. There are in fact two claims here. The first claim is that our experience is essentially an experience of temporality. Concerning the temporality of experience, we may notice that the project of Tina Röck is in good terms with a mainstream view in contemporary neuroscience concerning the essentially predictive function of the brain¹. This gives independent support to that claim.

To fully appreciate the temporality of experience, according to the author, we need to strip experience and our understanding of it of stable and atemporal conceptual influences and explanatory models by following the phenomenological procedure of epoché. This is a necessary condition for enabling the processual understanding of reality. A legitimate question to ask here is how we are going to understand anything without concepts. Such a perspective entails that a propositional processual understanding of reality would be impossible, because using words entails using concepts, unless we are engaged merely in empty chatter. Of course one may claim that understanding does not have to be

propositional; instead, it is only after understanding is achieved that we try to put it into words.

I find the second and general claim of the book more problematic. The claim concerns the following implication: in practice, we cannot escape from temporality; therefore reality is fundamentally temporal, changing and dynamic. We must keep in mind that this is not true *per se*. There is no guarantee that how we perceive the world is how the world actually is. Such a claim, however, seems to be justified within the framework of Husserl's phenomenology, and under the presumption of a unity of the internal and external. It seems that Röck's ultimate task is to merely establish a correlation between thought and reality based on adequacy. In other words, the co-relata of dynamic and changeable thought should also be dynamic and changeable. As she points out, Whitehead was looking at coherence, and so we are looking at coherent integration of thought and reality along the lines of dynamicity, and Husserl's phenomenology gives us a great clue about how to do it. However, we still have to keep in mind that coherence does not entail actual existence, and the standard term for this is "Foley rationality"². Hence, I think that many readers will find the general claim of the book problematic, because the above implication is invalid even if phenomenology entails dynamism in the world. Another related issue here is the author's trust that the intersubjective entails the objective. Well, it does not. We cannot rule out the possibility that we are entrapped in a massive hallucination. Hence, the two terms are not synonymous. In sum, the reader would like to see a discussion of the classical dichotomy between appearance and reality, which grounds all the skeptical arguments in philosophy.

In conclusion, the book makes important contributions to the field of process philosophy by embedding dynamic realism in the phenomenological tradition. It provides a convincing demonstration that the marriage between these two speculative paradigms promises to be happy. However, beyond that marriage there are serious problems that cannot be overlooked. Since they are not addressed in the book, I would say it is strictly for philosophers with narrow interests in these two traditions.

¹ See, for instance Barrett, Lisa Feldman (2017)

² Foley, Richard (1987)

References

- Barrett L F (2017) The theory of constructed emotion: an active inference account of interoception and categorization, *Soc Cogn Affect Neurosci.* Jan; 12(1): 1–23.
- Foley R (1987) *The Theory of Epistemic Rationality*, Cambridge: Harvard University Press.