Perspectives on the self

PERSPECTIVES ON THE SELF

Edited by
BORAN BERČIĆ

University of Rijeka

PERSPECTIVES ON
THE SELF

Edited by
BORAN BERČIĆ

University of Rijeka 2017

PERSPECTIVES ON THE SELF

Edited by
BORAN BERČIĆ

Preface and Acknowledgements

This collection contains seventeen articles on the self and related subjects. All are published here for the first time. The collection covers a wide range of topics: metaphysics, philosophy of mind, philosophy of science, philosophy of language, history of philosophy (modern and ancient, eastern and western), aesthetics and ethics. This variety explains the title - *Perspectives on the Self.*

The occasion for the volume was a conference on *The Self* held on March 31 and April 1 2016 at The Faculty of Humanities and Social Sciences in Rijeka, Croatia. I wish to thank to all those who participated in the conference and submitted their contributions for this collection. Also, I wish to thank to Eric T. Olson, Takashi Yagisawa, Luca Malatesti and Leonard Pektor for the language proofreading of the articles in the collection.

This collection is the end product of the activities of a group of philosophers from the Rijeka Department of Philosophy and colleagues who have worked with them. The activity of this group started in the autumn of 2010 as an informal weekly seminar on identity. Philosophers made up the core of the group, although colleagues from the departments of Psychology and Literature also took part. The main support for these activities was the research project *Identity* of the University of Rijeka (http://identitet.ffri.hr). Many of the articles in this collection are written as part of the work on this research project. We hereby express our gratitude for this support. It made possible the visits of the colleagues from other centers and countries. On several occasions Yagisawa, Olson, Kardaš and other colleagues visited Rijeka and worked with the group. Finally, it was the support that made publication of this collection possible.

BORAN BERČIĆ
May 2017

CONTENTS

CONTRIBUTORS

BORAN BERČIĆ
Department of Philosophy, University of Rijeka, Croatia
(boran.bercic@ri.t-com.hr)

MARIN BIONDIĆ
School of Electrotechnics Rijeka; Department of Philosophy, University of
Rijeka, Croatia
(marinbiondic@yahoo.com)

ZDENKA BRZOVIĆ
Department of Philosophy, University of Rijeka, Croatia
(zdenka.@uni.ri)

FILIP ČEČ
Department of Philosophy, University of Rijeka, Croatia
(fcec@ffri.hr)

ANA GAVRAN MILOŠ
Department of Philosophy, University of Rijeka, Croatia
(anag@ffri.hr)

LJUDEVIT HANŽEK
Department of Philosophy, University of Split, Croatia
(ljuhan@ffst.hr)

MARKO JURJAKO
Department of Philosophy, University of Rijeka, Croatia
(mjurjako@gmail.com)

GORAN KARDAŠ
Department of Philosophy, Department of Indology, University of Zagreb,
Croatia
(gkardas@yahoo.com)

LUCA MALATESTI
Department of Philosophy, University of Rijeka, Croatia
(lmalatesti@ffri.hr)

MILJANA MILOJEVIĆ
Department of Philosophy, University of Belgrade, Serbia
(miljana.milojevic@gmail.com)

Nenad Miščević
Department of Philosophy, University of Maribor, Slovenia; University of Rijeka, Croatia; CEU Budapest, Hungary
(vismiscevic@ceu.edu)

Arto Mutanen
Finnish National Defence University, Finnish Naval Academy, Finland
(arto.mutanen@gmail.com)

Eric T. Olson
Department of Philosophy, University of Sheffield, England, UK
(e.olson@sheffield.ac.uk)

Matej Sušnik
Department of Philosophy, University of Rijeka, Croatia
(msusnik@ffri.hr)

Márta Ujvári
Department of Philosophy, Corvinus University of Budapest, Hungary
(marta.ujvari@uni-corvinus.hu)

Iris Vidmar
Department of Philosophy, University of Rijeka, Croatia
(ividmar@ffri.hr)

Takashi Yagisawa
Department of Philosophy, California State University,
Northridge CA, USA
(takashi.yagisawa@csun.edu)

Introduction: Editor's Overview
BORAN BERČIĆ

Eric Olson in "The Central Dogma of Transhumanism" argues that we cannot upload ourselves into computers and continue our existence as cyber beings. Nick Bostrom and other transhumanists believe that this is in principle possible and that it is only a matter of current technological limitations that we cannot do so (the central dogma). However, Olson argues that this is in principle impossible (metaphysically impossible). He claims that we cannot be "sent as a message by telegraph or dictated over the phone" simply because we are material beings and "you cannot move a material thing from one place to another merely by transferring information." This is also the problem with *Star Trek* teleportation. If the process is understood not as a transfer of matter but rather as a transfer of information only, then the person who is assembled on board of the *Enterprise* cannot be numerically the same person as the one who was disassembled at the surface of a planet, but only its perfect replica. Olson explicates three presuppositions of the central dogma: "that there can be genuine artificial intelligence, ... that we can become computer people, ... and that technology can advance to the point where we could actually do these things." He is especially critical of the second presupposition. Interesting to note, the second presupposition faces the same problem as the idea of resurrection: How can we decay in our graves but nevertheless continue to exist somewhere else? Also, there are two more problems about the second presupposition: the branching problem and the duplication problem. If we could upload ourselves into a computer, then we could upload ourselves to several computers and continue our existence not as a single person but as several persons (the branching problem); and there would be no difference between the original person being uploaded into a computer and a new person being created in a computer (the duplication problem). To support the intuitions about the duplication problem, Olson puts forward a nice thought experiment with the British and Austrian Wittgenstein Societies. Both societies are in possession of a detailed scan of Wittgenstein's brain shortly before his death. The British Society decides to create a replica of Wittgenstein (they do not want to disturb a deceased person), while the Austrian Society de-

cides to recreate the original. Could there be any difference between the two? The branching problem and the duplication problem are seen as two sides of the same coin, so the question is whether the duplication problem has any weight of its own. Further, Olson compares three views about the metaphysics of human people: the pattern view, the constitution view, and the temporal-parts view. Transhumanists essentially rely on the assumption that we are patterns (Bostrom, Kurzwell, Dennett) and patterns can be transferred as information. Patterns can branch and duplicate. However, Olson argues that we are not patterns. We are particulars, not universals. We are things, not their properties. And this is why we cannot be uploaded into computers. (As we will see, Milojević argues that the self should not be understood as an entity but rather as a set of functions.) Olson also rejects the constitution view and the temporal-parts view, though he believes that the temporal-parts view is the most promising strategy for transhumanists. Due to the principles of arbitrary temporal parts and unrestricted composition, I can have a flesh-and-blood temporal part as well as a silicon-and-wire temporal part. Of course, these principles are highly problematic, but they provide a promising metaphysical framework for the transuhumanist idea that we can continue our existence in computers and on the internet. Although Olson finally rejects the temporal parts view, perhaps he is more permissive here than he should be. The principle of unrestricted composition does not allow us to combine temporal parts that belong to different ontological categories. We cannot be things (particulars) until t and patterns (universals) after t. That would be too much, even for the temporal parts view. Finally, Olson examines the option that transhumanist views, although metaphysically incorrect, can nevertheless be good enough for practical purposes. If uploading into a computer will give me everything that I could want of immortality, who cares whether metaphysical criteria of personal identity are satisfied or not? However, it seems that transhumanist ambitions cannot pass the practical concern test. We would not be concerned for computers filled with information about us in the same way and with the same intensity as we are concerned about ourselves.

Miljana Milojević in "Embodied and Extended Self" argues that we are essentially embodied but that we can also be extended beyond the limits of our bodies. Under special circumstances, certain artefacts or features of the environment can literally be parts of *us*. She argues that famous Otto's notebook is literally a part of *himself*. (Otto has Alzheimer's and cannot remember anything without his notebook.) Milojević wants to show that "the material body of the subject as well as some parts of his environment play a much greater role in the constitution of the self than is traditionally

thought." In order to support this claim she relies on several philosophical theories and assumptions. Four main ones are the following: (1) Functionalism in the philosophy of mind: she argues that the self should be seen as a set of functions, not as an entity of this or that kind, as immaterialists and animalists see it. (As we saw, Olson argues that we are entities, not patterns or sets of functions.) In the debate between role functionalism (mental states are identified with functions) and realizer functionalism (mental states are identified with typical realizers of these functions), Milojević rejects role functionalism and embraces realizer functionalism. "A realizer functional ontology of the self which takes into consideration bodily and environmental factors has the best chance of capturing all what is important for personal identity." This enables her to claim that (2) we are essentially embodied – that our cognition essentially depends on our bodily constitution and environmental factors. The idea is that our mind is constrained by our body. Here she relies on the insights of Gallagher, Shapiro, Noë, and others. However, some authors reject functionalism as incompatible with embodiment because of the multiple realizability of the mental (Shapiro). But Milojević argues that functionalism is compatible with embodiment. "Multiple realizability is not an enemy to embodiment, but only allows for different types of embodiment." Further, Milojević accepts (3) a psychological-continuity criterion of personal identity. Here she relies on Parfit's idea of overlapping chains, and particularly on the idea that narrative memory is essential for psychological continuity and therefore constitutive for personal identity (Wilson and Lenart). Finally she accepts (4) the extended-mind thesis, the view that our cognitive processes can be partly realized in devices external to our brains and bodies. "If we take a functionalist stance toward the mind, there are no *a priori* reasons for excluding non-neural matter from the realization base of mental properties." This does not mean that every device that we use is a part of our self. Two conditions have to be satisfied: the integration condition and the functional psychological condition. On these four grounds Milojević argues that Otto's notebook is literally part of him. Since Otto cannot sustain his narrative autobiographical memory without his notebook, his notebook is literally part of his self. In the same way, if we would literally not know who we are without our diaries and family photo albums (due to a certain kind of amnesia), then our diaries and family photo albums would literally be essential parts of our selves. It would be interesting to examine the consequences of switching the criterion of the ultimate self (a possible step Milojević does not talk about in her article). If we reject the criterion of narrative autobiographical memory and accept instead, say, a criterion of the physical and social impact that we have as agents, then our cellular

phones, laptops, cars, and bulldozers can become literally parts of *us*. This is, of course, assuming that we form functional wholes with these devices. Would this be an absurd consequence indicating a flaw somewhere in our reasoning, or perhaps an illuminating insight showing that we really are extended far beyond what we think?

Zdenka Brzović starts her "The Immunological Self" with a short list of the most plausible candidates for the identity criterion for a biological organism. However, it seems these candidates are not good enough and that we do not have a satisfactory criterion. *Functional integration* includes parts of an organism (cells) as well as groups of organisms (bee-swarms) or symbiotic organisms. Therefore, it is not satisfactory, at least not without further specifications. *Autonomy* relies on the insight that an organism is something that is able to sustain itself. However, it seems that "unicellular constituents of multicellular organisms" are also able to sustain themselves. *Genetics* cannot differentiate between identical twins, and has the counterintuitive consequence that acres and acres of mushrooms should count as a single organism. After this, Brzović focuses her analysis on the fourth proposed criterion - *Immunology*. Obviously, the very idea of immunology is closely related to the self. The immune system is a system with which an organism defends and sustains *itself*, it protects *itself* from harmful external influences. Our immune system distinguishes *us* from factors that are external to *us*, it "knows" whether it deals with *us* or with factors that are foreign to *us*. The *immunology criterion* has several versions. The oldest and the most striking is the *self-nonself theory* (Burnet). The self is "that which the organism's immune system tolerates (does not attack)." However, Brzović notes that this cannot be the criterion of the biological self. (Just to note, if this were the criterion of identity for an organism, then autoimmune diseases would be conceptually impossible.) The criterion must be some property that we have and that our immune system detects: our genes, our HLA tissue markers (molecular "identity card"), or some other property that *we* have and intruding organisms do not have. However, it seems that these criteria do not fit all the relevant facts (autoreactivity, pregnancy, transplantations, immune tolerance, intestinal bacteria, etc.). "All the phenomena examined demonstrate that it is not the case that the organism tolerates the self and rejects the nonself." Although generally critical about the proposed criterion, Brzović makes a concession in the case of autoimmune diseases: "autoimmune diseases are not considered as problematic since the self is defined by the immune system of the organism that is functioning properly." But when does the immune system of an organism function properly? Among other things, when it does not attack itself! But this is circular! So, autoimmune diseases

are not a problem for the immunological criterion only if normal functioning can be defined in a non-circular way. That is, without the assumption that the normal immunity system is one that does not attack the organism to which it belongs. But it is hard to believe that normal functioning can be defined without this assumption. Brzović concludes that talk of the self in the self-nonself theory can be taken only as a metaphor (Moulin, Tauber), not as an explicit identity criterion for organisms. In the rest of the article Brzović analyzes a few more versions of the immunity theory, so called *systemic theories of immunity*. In these theories the self is primarily seen as an *autopoetic* entity (Maturana and Varela, Jerne). However, "the main problem with views of this type is that they are vague so that it is not entirely clear what the main contribution consists in." The second version of the systemic theory is so called *danger theory* (Matzinger) "according to which the immune response is initiated by the fact that the immune system recognizes the substance as dangerous." Brzović objects that this theory does not have clear testable consequences. Of course, on the conceptual level the problem is that danger has to be danger for somebody. For this reason the danger theory cannot serve as a criterion for the identity of an organism because it presupposes it. Third version of the systemic theory is *continuity theory* (Pradeu), according to which the immune system reacts to patterns that differ from the ones it usually encounters. Brzović is sympathetic to the continuity theory because at least in principle it has clear testable consequences. However, she objects that this theory heavily relies on the functional integrity criterion, which is, as we saw, not clear enough. Brzović's conclusion is that all immunity theories of the self, if taken as a criterion of identity, have a fatal flow: they cannot serve as a criterion of identity because they presuppose it.

Nenad Miščević in "The Value of Self-Knowledge" draws a distinction between two main kinds of self-knowledge. The first kind is "knowledge of inner phenomenal states (that I feel pain in my back)." The second kind is "knowledge of one's causal and dispositional properties (that I am a gourmet or that I am prone to jealousy)." Miščević mentions other authors who draw analogous distinctions: between trivial and substantial self-knowledge (Cassam), or between first-personal and third-personal self-knowledge (Coliva). The first kind of knowledge is widely discussed in contemporary analytic philosophy, while the second was especially discussed by the ancients. Explaining the difference between these two kinds of self-knowledge, Miščević quotes Hatzimoysis, who said that "for the ancients self-knowledge is primarily a good to be achieved, whereas for the moderns it is mainly a puzzle to be resolved." However, in Miščević's view, the second kind of self-knowledge (knowledge of one's own causal

and dispositional properties) starts at a very basic level (Perry, Campbell, Damasio, Bermúdez). He illustrates the distinction with the following example: he sits at his desk and (1) he knows that he has a pain in his lower back, (2) he knows that the pain is related to his posture, and (3) he knows that the pain will stop if he straightens up. He straightens up and the pain stops. Of course, (1) is an instance of self-knowledge of the first kind, of inner phenomenal states. However, (2) and (3) are instances of self-knowledge of the second kind, of causal and dispositional properties. This might look surprising because (2) and (3) seem much closer to (1) than to the ancient *Know Thyself!* needed for the virtuous life and eudaimonia. However, since (2) and (3) are causal, Miščević categorizes them as cases of the second kind of self-knowledge, together with knowing that one is a gourmet or that one is prone to jealousy. After this, Miščević proceeds to the question of the value of self-knowledge. He accepts the usual distinction between extrinsic value (instrumental) and intrinsic value (in itself). These two distinctions yield a logical space of four options: (1) instrumental value of knowledge of inner phenomenal states, (2) intrinsic value of knowledge of inner phenomenal states, (3) instrumental value of knowledge of one's causal and dispositional properties, and (4) intrinsic value of knowledge of one's causal and dispositional properties. Some authors believe that knowledge about our own inner phenomenal states is trivial (Cassam). However, Miščević strongly rejects this view and argues that knowledge of our own inner phenomenal states is essential for our survival: without knowing that we are in pain, or thirsty, or hungry, ... we could literally not survive. Of course, the question here is whether I eat because I am hungry or because I know that I am hungry. It seems that our inner states move us and have instrumental value for our survival, not our knowledge of our inner states. Miščević supports his claim with the case of *analgesia*. But it is questionable whether analgesia really supports his point because analgesia is not a condition where we do not know that we feel pain, it is a condition where we simply do not feel pain. For this reason he argues that knowing that one is in pain just is being in pain (in this context he talks instead about awareness). Although some authors reject this identification (Coliva), Miščević insists on it. Further, Miščević argues that, besides enormous instrumental value, our knowledge of our own inner phenomenal states also has enormous intrinsic value. He argues that it is *constitutive* for us: "If the phenomenal light within were replaced by such a darkness, you would turn into a zombie, and stop being who you are." But here we face the same question again: the problem with zombies is not that they lack knowledge about their mental states, the problem is that they lack mental states. Therefore, Miščević's claim that knowledge about our own

inner phenomenal states has enormous intrinsic value because it is constitutive for us rests on the assumption that we have a mental life iff we know that we have it. Further, Miščević analyzes the value of knowledge about one's own causal and dispositional properties. He rejects the view that such knowledge "has no deeper value" (Feldman and Hazlett, Cassam). He also rejects the argument, or rather just intuition, that selfconscious Sam lacks authenticity that unselfconscious Sam has. In his opinion, unselfconscious Sam lacks something else – coherence. Here Miščević relies on Lehrer and claims that: "In order to live wisely one has to fulfill a first-level and a second-level condition: on the first level to have correct action-guiding preferences, and on the second level coherent reflective mechanisms." Miščević also analyzes famous literary characters that lack second-order insight into themselves: prince Myshkin from Dostoyevsky's *The Idiot* and Platon Karataev from Tolstoy's *War and Peace*. In his view, what we find admirable in such characters is not their lack of second-order insight, but rather their "so admirable first-order moral qualities that compensate for the lack of reflection." At the end of the article, Miščević wonders what is the relationship between the value of curiosity about *p* and the value of the answer to the question about *p*. Is our curiosity valuable because the answer is valuable, or is the answer valuable because our curiosity is valuable? What comes first? Miščević opts for the response-dependentist answer but leaves this discussion for another occasion. He concludes his article with the claim that *Know thyself!* "is still good advice after two and half thousand years."

Luca Malatesti in "The Self-ascription of Conscious Experiences" wants to find out how do we ascribe experience to ourselves. Paradigmatic cases are statements like "I experience pain in my elbow" and "I have an experience of red." He wants to know what one needs in order to make statements like these, that is, to ascribe experiences to oneself. First of all, we need concepts, and concepts are "ways of thinking about objects, properties and other entities." Malatesti starts his analysis with color perception and argues that having a corresponding experience is a necessary condition for having a concept. That is, he starts his analysis with so called *phenomenal concepts*. Relying on Jackson's knowledge argument (Mary), Malatesti rejects behaviorism, physicalism and functionalism in the philosophy of mind (Ryle, Smart, Putnam) and claims that: "The relevant concept of conscious experience is that unique concept *C* to possess which a thinker must meet the condition that she has had experience *e*." With concepts we form thoughts, and thoughts are "wholly communicable" (Dummett). Perhaps there is a certain tension here between subjective experience and intersubjective thought. Nevertheless, in parts 3. and 4. of this article Malatesti proceeds to the next step of his analysis, and this step is crucial. Whenever we see

that (1) the rose is red, in a sense we know that (2) we have experience that the rose is red. But the question is how we make the step from (1) to (2). How do we make the step from properties of the world to the properties of our experience? This step Malatesti calls *compelling transition* or *central transition*. Malatesti rejects a quasi-perceptual model of self-awareness that relies on the idea of an *inner sense* or *inner scanner* (Armstrong), because we cannot "formulate demonstrative thoughts" about our own experience (Shoemaker). Our own experience is not something that we are directly aware of. The second model of self-awareness that Malatesti discusses has its ground in the idea that our experience is *transparent* (Moore). Since a description of our experience of the world seems just the same as a description of the world, one might be tempted to conclude that the step from (1) to (2) is trivial and automatic. However, Malatesti rejects such a view. He says that "from the judgment that something is red, it cannot follow that I am having an experience of red." The observational concept $SQUARE_1$ need not be the same concept as $SQUARE_2$ that is used in inferential reasoning. An reasoner could not infer a priori that something is $SQUARE_2$ from the fact that it is $SQUARE_1$. Finally, in part 5. Malatesti says something about the concept of the self that we must have in order to be able to ascribe conscious experience to ourselves. Following Millar, he says: "The mastery of the concept of conscious experience involves the capacity to think about ourselves as entities that have sense organs and internal states that are determined by interactions with certain sorts of stimulation of these sense organs."

Boran Berčić in "The Logical Positivists on the Self" examines the views of logical positivists about the nature of the self (Schlick, Carnap, Ayer, Weinberg, Reichenbach). In the first part of the article author shortly compares four ways in which we can understand Descartes' *Cogito*: (1) as an expression of a nonpropositional immediate awareness of our own existence, (2) as a proposition, an *a priori* truth of reason, (3) as an inference, with or without underlying substance–attribute ontology, and (4) as a performance, true by uttering it. Although this is not decisive for the rest of the article, author accepts (3) in its ontological reading. He claims that *Cogito* should be understood as an inference from attribute to its substance. In the second part of the article author analyses logical positivists' critique of the Descartes' argument. (1) Schlick argued that *Cogito* is not a proposition at all, but rather a stipulation, or a concealed definition. (2) Carnap believed that *Cogito* is meaningless because it cannot be formulated in the language of logic. (3) Weinberg argued that *Cogito* could be understood as a valid inference, but then it would be a tautology and could not serve Descartes' purposes. (4) Ayer claimed that *Cogito* is an invalid inference,

an instance of *non sequitur*. After the critique of Descartes, where positivists said what self *is not*, author passes onto the positive part of their view where they say shat what self *is*. (1) Carnap argued that "self is the class of elementary experiences." He hoped that the concept of a class will help answer a standard objection that a self is not just a bundle of experiences. However, Berčić is skeptical about this solution: although concept of a class does express what elementary experiences have in common, it does not express the interconnectedness that elementary experiences should have in order to form a self, that is, in order to account for the unity of consciousness. Although Carnap's overall programme in the *Aufbau* is certainly reductionist, Berčić argues that, in a sense, Carnap was antireductionist about the self. (2) Ayer claimed that "self is a logical construction out of sense-experiences," where *X* is a logical construct out of *a, b, c, ...* iff sentences about *X* can be translated into sentences about *a, b, c, ...* . Of course, the question is whether such reduction can preserve all the facts about the first person perspective, but author does not enter into this problem. Ayer believed that he can solve some difficulties that Hume has faced, for instance, he argued that different sense-experiences belong to the same self because they are related to the same body. Ayer heavily criticized underlying assumptions of Cartesian philosophy of mind. As a positivist, Ayer accepted *neutral monism* and argued against Cartesian introspectionism. Berčić presents his argumentation as a tension between (i) *I* and *world* are constructed out of neutral elements, and (ii) I can doubt the existence of the whole world but I cannot doubt my own existence. Also, Ayer believed that body is essential in acquiring a concept of a self. Therefore, there is a tension also between (i) I can develop a concept of a self only if I have a body, and (ii) Once I develop a concept of a self, I can doubt whether I have a body. (3) Reichenbach argued that "Ego is an abstractum composed of concreta and illata," where *abstractum* should not be understood as abstract entity in a nowadays sense, as something "out of space and time," but rather just as a composite entity. We are composed of our body (concretum) and our mental states (illata). Reichenbach insisted on the point that our own mental states are *illata* or inferred entities, not something that is immediately given in the introspection. His critique of the Cartesian programme in the philosophy of mind can be summed up in five points: (i) Self is not something simple, it is something composed of elements. (ii) Self is not known by a direct insight, but indirectly and gradually. (iii) Self is not the Archimedean point of the knowledge, it is discovered later in the process of the rational reconstruction. (iv) Self is not known *a priori* but *a posteriori*, its existence is an empirical discovery. (v) Self is not something that exists necessarily, its existence is contingent. In the fourth part of the

article Berčić examines logical positivists' answer to the objection that reductionism about the self is circular because experience presupposes self. Positivists were well aware of this objection and they offered an elaborated answer: although we start with our own experience we do not know at the beginning that it is experience and that it is ours, we find it out later. In order to analyze this argumentation Berčić draws a distinction between three senses of reductionism: (1) conceptual, (2) epistemological, and (3) ontological. He argues that, although logical positivists were reductionists about the self in all three senses, their reductionism should primarily be understood as (2) epistemological reductionism. That is, as the claim that in order to know what self is, we have to know what its elements are.

Ljudevit Hanžek in "Brentano on Self-Consciousness" critically examines Franz Brentano's views from his *Psychology from an Empirical Standpoint* (1874), as well as views of several contemporary authors who have defended a Brentanian view about self-consciousness. In order to avoid an infinite regress of mental states, Brentano assumed that our mental states have a quality of inner consciousness. The idea is that whenever we are aware of an object, we are *ipso facto* aware that we are aware of that object. In other words, our awareness of our awareness is already contained in our awareness. The question is whether this idea can be worked out in a satisfactory way. Hanžek argues that it cannot. Besides Brentano's own views, Hanžek analyzes several similar proposals of contemporary authors and rejects them all. Uriah Kriegel relies on the distinction between focal and peripheral awareness. However, Kriegel's peripheral awareness cannot serve the purpose of Brentano's inner consciousness. Hanžek also argues that the usual distinction between transitive and intransitive consciousness (Kriegel, Gennaro, Rosenthal, Byrne) cannot help here. Intransitive consciousness cannot play the role of Brentano's inner consciousness. Finally, although Hanžek finds Amie Thomasson's interpretation of Brentano interesting, he rejects it as insufficiently supported by the textual evidence from Brentano's work. In several places in the article Brentano's view is expressed by saying that "a mental state is aware of itself" or similar formulations. But how can a single mental state be aware of itself? How can it be aware of anything? Only a cogniser as a whole can be aware of something, including its own mental state. Maybe "a mental state that is aware of itself" is just a clumsy way of saying something sound, but maybe that is just what Brentano had in mind.

Goran Kardaš in "The No-Self View in Buddhist Philosophy" presents and analyzes Buddhists' arguments for their claim that there is no such thing as the self. Generally speaking, Buddhists were empiricists who criticized metaphysics. They were eliminativists or reductionists about the

self and criticized antireductionists who argued that the self is an entity that exists on its own. A general argument that Kardaš analyzes is directed against earlier metaphysicians in the Indian tradition who believed in a one-to-one correspondence between language and reality (*nama-rupa*). "Who knows a name of a thing knows at the same time a thing itself referred to." Buddha was a conventionalist about language and rejected this idea. His argument is that from the fact that we think and talk about "I" (*aham*) and "self" (*atman*) it does not follow that "there exists a corresponding mysterious and undying entity called Self." Here metaphysicians are accused of the fallacy of substantivization or reification. A second general argument (that Kardaš only briefly states in 6.1.) is that the Self is supposed to be something permanent and not subject to any change. However, since nothing is permanent and everything is subject to change, such an entity as the Self simply cannot exist. A more specific argument is directed against the metaphysicians of the *Nyaya school*. They accepted a substance-attribute (*dravya-guna*) ontology and, like Descartes many years later, argued that since "pain, joy, knowledge, etc." are obviously attributes, there must exist a substance to which they belong, that is, the Self. On this picture the Self is an inferred entity, and experiences are "inferential marks of the Self" (*atmano lingam*). However, Buddha was not impressed with this argument: "Buddha is wondering, if we somehow could remove all cognitions, emotions, perceptions, volitions, etc. from our experience, would there remain anything that is the substratum of these properties?" We can guess that Buddha would deny that *Avicenna's floating man* (deprived of all sensory stimuli) would be aware of his self. After Buddha, his followers in the Abhidharma school defended *a bundle theory of the Self*. They argued that "there are foundational properties (*dharma*) of experience but not property-possessors (*dharmin*)." A second specific argument for the No-Self View is what Kardaš calls *Buddha's linguistic turn*. Buddha believed that the way in which we think and talk about experience can and should be depersonalized. He argued that we should not ask *Who craves?* but rather *What causes cravings?* "I feel pain" should be analyzed as "conditioned by *x*, *y*, ... (a feeling of) pain arises." (In the contemporary discussion about free will, this argument is called *The Disappearing Agent Objection*, though Buddha did not think it was an objection.) Kardaš claims that, even if we accept this argument, there is still a sense in which we can say that the Self exists: "Appropriating also a later Buddhist terminology, we can say that the concept of "self" (*atman*) is a cognitive construction (*vikapla*) or imputation (*samaropa*) formed on the basis of the stream of psychological events or "the stream of (causal) happening/becoming (*bhavasota*)."

Ana Gavran Miloš in "The Self in Ancient Philosophy" wonders how the ancients understood the self. She analyzes two opposed views on the matter. On the one hand, there are authors who argue that the ancient conception of the self was essentially different from the modern one, that it was not "subjective-individualistic," and that "Greeks never adopted a first-personal point of view" (Gill). Roughly speaking, the claim is that the ancient Greeks did not have a concept of self, at least not in the sense that we have it today, after Descartes' epistemology and Kant's ethics. According to this view, the Ancients did not have the idea of *subjectivity*. On the other hand, there are authors who argue that the ancients did have several concepts sufficiently similar to the modern concept of the self, and that therefore there is no essential difference between the way that we understand ourselves today and the way that the ancients understood themselves (Long, Sorabji). According to this view, the ancients, of course, did not have the modern Cartesian concept of the self as a source of epistemological certainty and privileged access (Burnyeat), but they did have the idea of "an individual owner who sees himself or herself as *me* and *me again*" (Sorabji). According to this view, the Ancients understood themselves as having both *objectivity* and *subjectivity*. Gavran Miloš argues in favour of the second option and wants to show that the Ancients did have an explicit or at least an implicit idea of *subjectivity* and first person perspective. The ancient notion that includes our notion of the self is the notion of the *soul (psyche)*. Therefore, she shortly analyzes views that Plato, Aristotle, and Epicurus had on the human soul (dualism, hylomorphism, and materialism). Her point is that, although they did not have the Cartesian idea of the self as epistemological rock bottom, they said a lot about the *ontological self* (What kind of thing am I?) and about the *ethical self* (How should I live?). On these grounds, contrary to Gill, Gavran Miloš claims that "the objective human self does not exclude an individual aspect of the self in ancient philosophy." She supports her claim with the quotation from Plato's *Phaedo* where Socrates talks about his immortal soul and says that "provided you can catch *me* and *I* do not escape you." In her opinion, *subjectivity* is indispensable for ethical reasoning because "the teleological-eudaimonistic framework of the self necessarily involves both an objective and a subjective aspect." Therefore, although there are some differences, it would be wrong to think that the ancients understood themselves in a way that was essentially different from the way that we understand ourselves today.

Matej Sušnik in "Ideal Self in Non-Ideal Circumstances" wants to unveil the nature of the relationship between the real and the ideal self. His starting point is internalism about reasons, the view that one's reasons for acting must be somehow grounded in one's actual motivation (Hume,

Williams). There is a strong motive for this view: reasons are supposed to move us, and it is not clear how could they move us if they were not grounded in the motives that we actually have. However, reasons for acting cannot just amount to actual motives because we are not always completely informed, rational, calm, disinterested, etc. For this reason the internalist has to *idealize* our actual selves and our actual motives. After all, reasons are essentially normative. Therefore internalists usually claim that "one's reasons are not dependent on the motivation of one's actual self, but rather on the motivation of one's ideal self." Now, the question is what is the relationship between us and our ideal selves, and how thinking about our ideal selves can help us in deciding what to do. Sušnik analyzes three answers, rejecting first two and accepting the third. (1) According to the straightforward model, I have a reason to do *x* in circumstances C iff my ideal self would do *x* in circumstances C. However, this model faces a problem because ideal self has motives that differ from the motives of the actual one (Johnson, Sobel, Smith, Markovits, Wiland). If I should better leave the room because I am upset, my ideal self would stay in the room because he is calm; if I believe that I am James Bond I should see a doctor, but my ideal self should not see a doctor because he does not have such a belief; etc. (2) According to the advice model, I have a reason to do *x* in circumstances C iff my ideal self would advise me to do *x* in circumstances C (Smith). This model seems to be better because my ideal self would tell me to leave the room and to see a doctor. However, this option faces a related problem: it is not clear how the advisor's motives are related to my actual motives. In other words, it is not clear in what sense my ideal self is *my* ideal self (Johnson, Sobel). Any reasonable person would tell me to leave the room and to see a doctor, it does not have to by *my* ideal self. "His identity is not important." And this is a serious problem for internalism because its central tenet is that the advice of my ideal self has to be somehow related to my actual motives. For this reason the advice model departs from the spirit of internalism. (3) According to the third model, I have a reason to do *x* only if there is a "*sound deliberative route*" from my actual motives to my doing *x* (Williams). Within this model my actual self must have access to my ideal self. It must be possible for me, as I actually am, to reach the viewpoint of my ideal self. Otherwise decisions of my ideal self cannot be relevant for me as I actually am. This is the model that Sušnik accepts. He believes that "we learn something about ourselves when we engage in the process of idealization ... what we really desire, what we plan to do, and what is the best way for us to proceed in given circumstances." Sušnik also discusses a closely related problem from ethics: What do we exactly have in mind when we talk about stepping into someone's shoes? If I say "If I

were you I would do *x*!" whose values and preferences do I have in mind, mine or yours? (Hare, Taylor) Sušnik believes that Williams' solution helps here as well. Although by stepping into other people's shoes we learn about them, "there is no point in imagining oneself in the shoes of someone else if that process implies that the agent needs to *become* someone else."

Filip Čeč in "The Disappearing Agent" analyzes and evaluates the strength of this argument (Pereboom) in the context of the contemporary debate about free will. Contemporary libertarianism has two main versions: agent causal libertarianism and event causal libertarianism. Agent causal libertarianism is the view that agents are causes of their actions, that our actions are caused by *us*, that *we* are causes of our actions (Chisholm, O'Connor, Clarke, Griffith, Steward). Although this might seem completely plausible at a first glance, it seems that it implies a weird picture of ourselves. What kind of things are we who cause our actions? Kantian noumenal selves, Aristotelian unmoved movers? After all, how could a substance cause anything? It seems that agent causal libertarian is committed to an antireductionistic and therefore ontologically problematic understanding of the self. "The notion of causation invoked by the agent-causalist is not reducible to causation among events, ... rather ... it invokes an ontologically specific kind of selfhood ... which is irreducible to event ontology." For this reason event causal libertarians "opted for an ontological framework based exclusively on states and events." This framework can contain "states and events involving the agent" like desires and beliefs, but cannot contain selves or agents (Kane, Ekstrom, Balaguer, Franklin). In addition to this, event causal libertarians understand free will as something essentially indeterministic. In their view, the paradigmatic cases of free decisions are so-called torn decisions (Kane, Balaguer, Franklin). "The paradigmatic notion of libertarian event-causal decision making is exemplified in various instances of torn decision making" (Kane). Torn decisions are cases where we have equally strong reasons for two options and some indeterministic event makes us choose one option instead of another. Čeč analyzes a notion of torn decision in detail, and offers a list of six conditions a decision has to satisfy in order to be torn. One might say that torn decisions are cases where Buridan's ass tosses a coin. (Also, torn decisions are supposed to be character building, but that is beside the point in this context.) Of course, the question is how an action that is by definition a result of a pure chance can be *free*, and how it can be *mine*. If it is a result of chance, then it cannot be something that I did, it rather has to be something that happened to me. (Additional problem is that determined actions also cannot be mine because they are determined.) If all my free actions are caused by chance, then I cannot be an agent since I do not cause anything. This is the

disappearing agent objection, and it is put forward as an argument against event causal libertarianism, usually by agent causal libertarians. However, here Čeč relies on a *tu quoque* strategy and argues that the disappearing agent objection is a problem for agent causal libertarianism as well. "It seems strange to say that the situation of motivational equipoise should be resolved by the agent." Really, it is not clear how Buridan's ass would do any better with a noumenal self than without it. Equally strong reasons are equally strong reasons, whether they are realized in a neurological basis or in an immaterial and eternal soul. Čeč defends event causal libertarianism and discusses five possible ways that an event causal libertarian might react to the disappearing agent objection. For instance, one might try to offer a reductionistic account of the self: "invoke a notion of plural voluntary control of the agent over his options" (Kane); "use the notion of appropriate non-randomness" (Balaguer), argue that the agent identifies with one option (Velleman); rely on the phenomenology of decision making; etc. However, although he is inclined toward Balaguer's solution, Čeč argues that none of these options is completely satisfactory, and that the event causal libertarian has to accept "a bit of residual arbitrariness in his ontology." He claims that, in spite of this arbitrariness, agent will not disappear. It seems that disappearing agent objection has even wider relevance. There is something horrifying in determinism. Its implication that all our future decisions are determined certainly causes some anxiety, but what is really horrifying is its prospect that we as agents do not exist. *We are illusion, we do not really exist!* is the insight of the ultimate abyss. Though Buddha believed that this insight is in fact a relief (see Kardaš' article in this collection).

Marko Jurjako in "Agency and Reductionism about the Self" explores the question whether the psychological criterion of personal identity (Parfit) is compatible with the agency based account of the self (Korsgaard, Bratman). He argues that it is. Since agency necessarily includes mental activity like desiring, intending, planning, etc, the psychological criterion, in some very broad and unspecified sense, obviously accommodates agency as well (Davis, Baker). Jurjako claims that although Parfit in his writings does not explicitly analyze agency, he does not rule it out either. However, the main problem for the compatibility of the two views is that psychological connectedness comes in degrees (Parfit), while it seems that agency does not (Korsgaard, Schechtman). Our memories can exist without unity but we as agents cannot. Parfit believes that we are mereological sums like nations, while Korsgaard believes that we are rather like states because we have an organizational principle. In other words, the psychological criterion is compatible with reductionism about the self, while the

agency based account is not. Jurjako rejects this conclusion and argues that reductionism about the self is compatible with the agency based account of the self. He explores the thesis that agents can be scattered through space and time just as memories are. To illustrate his point he proposes several related thought experiments. Here is one he puts forward at the end of the article: in order to escape the law, criminal X splits into Y and Z; after that Y and Z cooperate to carry out the original plan of X. Jurjako argues that in this case "Y and Z would be the same agent albeit spatially distributed." However, even if we accept the intuition that in some sense Y and Z would be the same agent, the question is whether this intuition supports the claim that reductionism about the self is compatible with the agency based view of the self. Although there is no unity of consciousness between Y and Z, what makes them the same agent is the fact that they stick to the original plan of X, who had a unity of consciousness at the time he made the plan. However, we can say in the same sense that construction workers are the same agent when they stick to the plan of the engineer, even though they do not have psychological continuity with the engineer as Y and X have with X. Another thought experiment that Jurjako analyzes is the following: imagine that X committed a crime and that after that, in a Parfit-like manner, he split into Y and Z. Are Y and Z identical to X? No! Are Y and Z guilty of the crime X committed? Yes! Jurjako believes that here we should introduce a distinction between *moral selves* and *selves of personal identity*. The difference between the two "consists in the fact that while the latter is unique to a person, the former comprises a set of mental states, personality traits, dispositions, and a history that, in principle, might be shared by different persons." For this reason, Jurjako argues, Y and Z should be punished for the crime X committed even though they are not identical to X. Of course, it is questionable what the intuitions here really are. What is meant by guilt and responsibility here? Perhaps we feel that society should be protected against people like Y and Z, or that they should be reformed, or that each of them should serve half the sentence, etc. After all, we do not sentence people for having the same personality traits as criminals; we sentence them for actually committing a crime. Generally speaking, the agency based account of the self certainly is reductionistic in a sense that it does not rely on Cartesian egos, immortal souls, or any other strange metaphysical entities.

Marin Biondić in "On Never Been Born" wonders whether we can talk about the people who have never been born. The old dictum that the luckiest people are those who have never been born is in fact very puzzling. For how we can say anything about the people who have never existed? To whom are we referring? We can meaningfully talk about people who

have died, but can we meaningfully talk about people who never came into existence? Biondić compares the views of several contemporary authors who have discussed the matter (Parfit, Benatar, Yourgrau). Biondić sides with Parfit and argues that we cannot meaningfully talk about people not yet born or feel sorry about the misfortune of those who are never born. "Nobody waits, in the waiting room of prenatal nonexistence, for his order to exist." The interesting consequence of this common-sense view seems to be that we should not feel any special gratitude for our existence to our creators (parents or God). Biondić also accepts Parfit's view that the evaluation of existence is a special case of evaluation: although it is good for us that we exist, we would not be worse off if we didn't. This might sound contradictory but it is a consequence of the view that we can evaluate only lives of actual people. The concluding General schema 4 might seem misleading because it suggests that there are two sorts of non-existent people: those who never exist and those who do not yet exist. Perhaps there is a sense in which actual people were not-yet-existent before they were born, but it is not clear how there could be any sense in which we could talk about never-existing people.

Iris Vidmar in "Fictional Characters" compares two approaches to the nature of fictional characters: the approach of logicians, metaphysicians and semanticists - LMP approach, and the approach of literary aesthetics - LA approach. In the LMP approach people discuss "questions of reference and denotation, truth conditions, and meaning of nonexistent objects or abstract entities," while in the LA approach they focus on "the way fictional characters come to life within the established literary practices (including, roughly, writing, reading and discussing literary works)." Vidmar argues that for the right understanding of the nature of fictional characters we should primarily focus on the LA approach, but, since her proposal is syncretic in nature, she claims that we should not neglect the LMP approach. Vidmar believes that her proposal is akin to Amie Thomasson's artifactualist theory of fictional characters. Further, Vidmar draws a distinction between internal and external perspectives on works of art. From the internal perspective we view fictional characters as real people in the real world; we think and talk about their motives, achievements, character traits, etc. On the other hand, from the external perspective we view fictional characters as fictional characters: we think and talk about the role they have in a novel, meanings they might have in relation to other works of art or cultural epochs. The fact that Emma Bovary might be seen as a fallen romantic hero is an external fact about her; the fact that the Blind Beggar "symbolizes and reinforces the blindness of every other character" in *Madame Bovary* is an external fact about him, etc. Consequently, we should distinguish between

the internal and the external identity of a fictional character, and both are constitutive of its overall identity. Fictional characters are composed of elements picked out of the real world and they can be seen as "place holders for the things that can happen to us." This is why we can emotionally engage with them. Vidmar also argues that the identity of fictional characters is relational since it is constituted by the active role of a recipient. The implausible consequence of this view is that there is no single Emma Bovary but rather as many Emmas as there are readers. Vidmar believes that this consequence is not as devastating as it might seem at first glance.

Márta Ujvári in "Haecceity Today and with Duns Scotus: Property or Entity?" analyzes the historical understanding of haecceity as an entity and the contemporary understanding of haecceity as a property, though the onus of her work is on the contemporary understanding. "The main role of haecceity in contemporary metaphysics is to secure the transworld identity of concrete individuals in non-qualitative terms." The main motive for positing haecceity is the fact that Leibniz's principle of the Identity of Indiscernibles fails to account for numerical identity. However, Ujvari warns us that this failure does not show that any qualitative account of numerical identity has to be wrong. A possible alternative is the neo-Aristotelian position where individual natures bear transworld identity (Fine, Gorman, Oderberg, Lowe). Contemporary authors often understand haecceity as a "relational property of being identical with itself" (Rosenkrantz, Diekemper). If haecceity is understood as a property then obviously it has to be a nonqualitative property. But what is a nonqualitative property? Diekemper answered this question by relying on the distinction between pure and impure properties (Adams, Armstrong, Loux). However, Ujvari rejects Diekemper's analysis and argues that he conflates impure qualitative properties with nonqualitative properties. Ujvari sides with Chisolm who argued that a property cannot "be conceived only by reference to a contingent thing." She rejects as inconsistent Rosenkrantz's claim that "Although an entity's haecceity is a relational property, an entity's intrinsic nature includes its haecceity." The traditional entity view and the modern property view are consequences of different motives and different ontological frameworks: "Today it is the Fregean function-argument of first order metaphysics, with Scotus it is the Aristotelian substance-accident framework." Since these two views are obviously incompatible, Ujvari believes that "there remains the task to find the proper ontological category for haecceity once its functional roles have been identified." Finally, Ujvari analyzes Gracia's instantiation-based approach to individuality. The main idea is that "individuality needs to be understood primarily in terms of the primitive notion of noninstantiability." She rejects Gracia's approach as in-

capable of accounting for genuine individuality. Instantiability can account for the difference between *F* and particular instances of *F*, but it cannot account for the difference between different instances of *F*, and this is exactly what haecceity is supposed to do. Although Gracia himself is aware of this problem, he does not offer a satisfactory solution. The moral here is that one should not conflate particularity with individuality. "Any sound theory of individuals, among other things of Selves, has to account for the feature of genuine individuality."

Arto Mutanen in "Who am I?" analyzes this question. He argues that it is "not a single question but a cluster-question to which different kinds of answers are expected" and that "different people are looking for different kinds of answers." He quotes Nietzsche's views on this question from *Ecce Homo* and *On the Genealogy of Morality*, and also Sartre's. Mutanen argues that the question "Who am I?" is a question of *identification*, where identification is primarily just a matter of determining who somebody is. "We ask who-questions if we do not know who somebody is. These questions are seeking information that allows us to identify the person." Here Mutanen quotes *Encyclopedia of the Social Sciences* from the 1930s, where identification is characterized "as dealing with fingerprinting and other techniques of criminal investigation" (Gleason). Mutanen insists on the distinction between *identity* and *identification*. "The question of identification is easily confused with the question of identity ... identification is a methodological notion and identity an ontological notion" (Gleason, Quine). He claims that "in philosophy, identity has been separated from identification, but in sociology and psychology such separation has not been done systematically." We may say that for Mutanen questions about identity are a matter of ontology (What is?), while questions about identification are a matter of epistemology (Who is?). Since "questions about identity look at the ontological characterization of what entity is," Descartes' dualism is a paradigmatic case of an answer to the question "Who am I?," if it is understood as a question about *identity*, and not as a question about identification. Descartes' point is ontological, he tells us what kind of entities we are. On the other hand, question about *identification* (about determining who somebody is) could be understood as "a question about locating oneself in society" (Gleason). Also, it could be understood as something that helps people to "feel that their life is meaningful – my membership of society is acknowledged: I know who I am." Further, author argues that identification is a modal notion and that possible world semantics is the appropriate framework for its understanding. Identification is sensitive to the opacity of context: Watson may know that Mr.Hyde is a murderer but not know that Dr.Jekyll is a murderer. There are possible worlds in which Dr.Jekyll

is not Mr. Hyde, where possible worlds are "worlds that characterize the knowledge Watson has." In his analysis Mutanen relies on the works of Hintikka, Quine, and Kripke.

Takashi Yagisawa in "Meta-Representational *Me*" analyzes first person singular *me*. He wants to show that *me* plays a fundamental role in the philosophy of language and in philosophy in general. Yagisawa starts his analysis by claiming that *me* and *self* are different notions. "The notion *me* applies to me and me alone absolutely, whereas the notion *self* applies to me relative to me, applies to you relative to you, ... Everyone is the self relative to her/him; ... But only I am me, period." Here one might object that the notion of *me* is reducible to the notion of *self*. However, Yagisawa rejects this objection arguing that something can be my self only if it is self to *me*, where *me* is primitive and it cannot be defined away. But what about *I*? Doesn't *I* ultimately amount to the same as *me*? Yagisawa accepts standard Kaplanian indexical theory of *I*, but claims that it "is not quite sufficient for giving a fully satisfactory explication of the notion *me*." Although it is a very good theory, Yagisawa argues, it "clearly fails to capture the uniqueness of the notion *me*." Also, he claims, standard indexical theory cannot explain the rigidity of "I." The model that is developed by Kripke for names and natural kind terms does not fit "I"; causation cannot play the same role in the case of "I" as it plays in the case of "tiger" or "Aristotle." Of course, we might think that "me," "myself," and "I" form a family of mutually definable terms that can be used interchangeably and are all equally basic and rigid. But, as we saw, Yagisawa disagrees and claims that "me" is basic and that only "me" assures the rigidity of other related expressions. Further, he argues that "What is essential to the notion *me* is not any notion of linguistic act but the notion of cognitive act, i.e., act of entertaining a content." It seems that Yagisawa here assumes that representations intrinsically contain *me*-way. "The content of my perception is put forth in the *me*-way, or *me-ly*." (He draws analogy with Chisholm's adverbial theory of perception.) Of course, here one might object that our experience simply does not contain such a thing as *me*-way or *me-ly*. Our experience of the world is our experience *of the world*, not of the *way in which the world is given to us*. The idea that there is such a thing as the way in which the world is given to us is not a part of the phenomenology of our experience, it is a false and misleading philosophical assumption. Yagisawa disagrees and rejects this objection. He further argues that the notion of "me" is based upon the "*me*-way" or "*me-ly*" of my perception, not the other way around. "The *me*-way does successfully lead me to the notion *me*, hence the postulation of myself as an entity." However, this claim is questionable: How can I know that the way in which I see the world is the way that I see it before I know

that I exist? The *me*-way cannot be the Archimedean point of epistemology, it is rather the result of the epistemological reflection. (Berčić defends the view of Reichenbach and Carnap who argued that only after substantial epistemological reflection we can know that our experience is *our* and that it is *experience* at all.) In order to justify the shift from "*me*-way" to "*me*," Yagisawa offers ontological analysis of the "Way-to-Thing-Shift." He offers examples of dancing a waltz, constellation of Orion, and curve ball in baseball. "Surely, a curve ball is a thing." In the part 7 of the article Yagisawa explains in detail how "*me*-way" assures rigidity. He argues that "The rigidity effect kicks in only when the *me*-way of representation gives rise to the first-person conception of the recipient of the representation as a result of the way-to-thing shift."

Part I

SELF AND BODY

1. The Central Dogma Of Transhumanism

ERIC T. OLSON

1. The Central Dogma

Transhumanism is a movement aimed at enhancing and lengthening our lives by means of futuristic technology. The name derives from the ultimate goal of freeing us from the limitations imposed by our humanity. Human beings are subject to many ills: disability, exhaustion, hunger, injury, disease, ageing, and death, among others. They set a limit to the length and quality of our lives. There's only so much you can do to make a human being better off, simply because of what it is to be human. But if we could cease to be human in the biological sense–better yet, if we could cease to be biological at all–these limitations could be overcome. An inorganic person would not be subject to exhaustion, disease, ageing, or death. The length and quality of her life could be extended more or less indefinitely. So it would be a great benefit, transhumanists say, if we could make ourselves inorganic.

They hope to achieve this by a process they call "uploading." The information in your brain is to be transferred to an electronic digital computer. The process does not merely store the information on the computer, as when you upload a letter of reference to a distant server, but uses it to create a person there: a being psychologically just like you, or at any rate a great deal like you. This person may be psychologically human, but not biologically. He or she would not be made of flesh and blood.

The aim is not merely to create new people in computers, but for *us* to move from our human bodies to the digital realm. The thinking is that the person created by the uploading process would be psychologically continuous with you: her mental properties would resemble and be caused by yours in much the same way that the mental properties you have now resemble and are caused by those you had yesterday. Given the widely held assumption that this is what it *is* for a person to continue existing–that personal identity over time consists in psychological continuity–the person in the computer would be you.

And once you are in or on a computer, you needn't worry about disease or injury or ageing or death. If the computer hardware that houses you is damaged, you need only move electronically to another piece of hardware. Travel would be as easy as emailing. You would not need food or shelter or furniture. The limitations imposed by human biology, or indeed any biology, would be a thing of the past. Your intelligence, patience, capacity for pleasure, and physical strength and stamina (if you are given a robotic body) could be enhanced indefinitely.

These hopes are founded on the extravagant assumption that the technology of tomorrow will literally make it possible to transfer a person from a human organism to a computer. Call this the *central dogma* of transhumanism. (The name is not meant to be pejorative; think of the central dogma of molecular biology.) The leading transhumanist Nick Bostrom puts it like this:

> If we could scan the synaptic matrix of a human brain and simulate it on a computer then it would be possible for us to migrate from our biological embodiments to a purely digital substrate (given certain philosophical assumptions about the nature of consciousness and personal identity). (Bostrom 2001)

Bostrom and others are confident that that we *could* "scan the synaptic matrix of a human brain and simulate it on a computer," and thus that such "migration" is possible.

The central dogma is of more than merely theoretical importance. If it really were possible for us to move from our human bodies to electronic computers, subject only to limitations of technology, it would mean that we are not doomed to wither and die. We are at least potentially immortal.

The central dogma raises many large questions. One is whether a "post-human" life would be as attractive and worthwhile as transhumanists imagine. Another is whether any of this is likely ever to happen. This paper is about the worries Bostrom puts in parentheses: whether it is metaphysically possible.

2. The Dogma's Presuppositions

The central dogma presupposes three contentious claims. The first is that there can be genuine artificial intelligence: it is possible for a computer not only to simulate intelligence and consciousness, but actually to *be* intelligent and conscious. More precisely, a computer could have the mental properties that you and I have. This will of course include those that make something a person, as opposed to a being with mental properties that fall short of those reqquired fot personhood in the way that, for instance, those of dogs do: such properties as self-consciousness. So it must be possible to

create a person just by programming a computer in the right way (and perhaps also providing appropriate connections to the environment). In other words, an electronic computer could be a person. Or perhaps we should say not that a computer could actually *be* a person, or be conscious and intelligent, but rather, more vaguely, that it could "realize" or "implement" a person or a conscious and intelligent being. (I will return to this point in a moment.) Call such a being a *computer person*. So the first presupposition of the central dogma is that there could be a computer person. This is what Bostrom means by "the assumption about the nature of consciousness."[1] I will call it the *AI assumption*.

The second presupposition is that you and I could *become* computer people. This is what Bostrom means by "the assumption about personal identity." It presupposes the AI assumption but does not follow from it. If I could become a computer person, then computer people must be possible; but the mere possibility of computer people does not imply that we ourselves could become such people. By analogy, it may be that there could be gods–conscious, intelligent beings who are immaterial and supernatural– even if it is metaphysically impossible for us to become gods.

In this regard the central dogma is like the doctrine of the resurrection of the dead: the claim that when we die and our physical remains decay, we do not perish, but continue existing in a conscious state in the next world–a place spatially or temporally unrelated to this one. This presupposes that there *is* a next world, some of whose inhabitants are people psychologically like us. But the mere existence of such a place would not make it possible for someone to get there from here. How could it be that I am totally destroyed in the grave, yet at the same time continue to exist with my psychology intact in the next world? That is the metaphysical obstacle to resurrection (van Inwagen 1978, Olson 2015). Transhumanism faces an analogous obstacle: how could it be that I am totally destroyed in the grave, yet continue to exist with my psychology intact in a computer? How is the "uploading" procedure supposed to bring this about?

The personal-identity assumption has an immediate and important implication, namely that uploading would not transform the computer itself– the physical object made of metal and silicon and plastic–from a nonperson to a person. This is because (according to the assumption) the person who ends up in the computer was previously in a human organism. She was not previously in the computer as a nonperson. It is the human person

---

[1] In calling it an assumption about the nature of consciousness rather than about the nature of the mental in general, Bostrom is presumably taking it to be uncontroversial that computers could have mental properties that do not require consciousness. This is doubtful, but I won't press it.

who becomes a computer person, rather than the previously unintelligent computer becoming a computer person. This implies that no computer could ever be a person itself. If a computer *could* ever be a person, or be conscious and intelligent, it could be made so by uploading–that is, by programming it in the right way. But in that case uploading would create *two* people or conscious beings: the former human person and the former unintelligent computer. The two computer people would be psychologically indistinguishable. Both would seem to remember my embodied past, one correctly and one falsely. How could either of them ever know which one he is? I take that to be absurd. So the personal-identity assumption entails that no computer could be conscious or intelligent. At best a conscious, intelligent being might "inhabit" or "be implemented on" a computer. (I will return to the question of what this "inhabiting" relation might be.)

The third presupposition of the central dogma is that it is possible for technology to advance to the point where we could actually do these things. This presupposes the first two claims, but does not follow from them. Even if uploading a human person into a computer is metaphysically possible, it may remain beyond any possible human capability. We might compare it with the task of creating a perfect physical duplicate of a human being. This is metaphysically possible: God could do it. But it's doubtful (to put it mildly) whether it could ever be possible for *us* to do it. Uploading might be like that.

I see no reason to feel hopeful about this third assumption, even if the others are true. But my interest is in the metaphysical assumptions, especially the one about personal identity.

3. The Branching Problem

Suppose for the sake of argument that the AI assumption is true: it is possible to make a digital computer into a person–or rather, to get it to "implement" or "realize" a person–by programming it in the right way. Even so, could a human person literally move to a computer? Transhumanists have had little to say about this. Some have defended the AI assumption at length (Chalmers 2010), but once they have established to their satisfaction that a person could exist in or on a computer, they have seen little reason to doubt whether we ourselves could do so. I think there are strong reasons for doubting it.

Here is one obvious worry. If someone could be uploaded into a computer, then someone could be uploaded into two computers. That is, the relevant information could be read off the human brain and copied simultaneously to two separate and independent pieces of computer hardware in just the way that transhumanists envisage its being copied to one. The

result would be two computer people, each psychologically just like the original human person. Each would have got his or her mental properties from the original person in the same way. So nothing could explain why one but not the other was the original person. More strongly, it seems that nothing could make it the case that one but not the other was the original person. If one were the original person, both would be. But they couldn't both be. There are two computer people in the story and only one human person, and one thing cannot be two things. If the original person and the first computer person are one, and the original person and the second computer person are one, then the first computer person and the second computer person would have to be one. (If $x=y$ and $x=z$, then $y=z$.) But they're not.

It appears to follow that a person could not move from a human body to a computer in the "double-upload" case. And if it's not possible in the double-upload case, it could hardly be possible in the "single-upload" case commonly imagined, because the same thing happens in both: the same information from the person's brain is transferred to a computer in the same way. So no amount of uploading is sufficient to make a human person into a computer person, contrary to the personal-identity assumption. Call this the *branching problem*.

The branching problem is familiar to anyone acquainted with philosophical discussions of personal identity. The reason is that it arises on almost any version of the psychological-continuity view–any view to the effect that an earlier person is the same as a later person just if the later person is in some way psychologically continuous, at the later time, with the earlier person as she is at the earlier time. (Psychological continuity is defined in terms of causal dependence of later mental states on earlier ones; for details see Shoemaker 1984: 90.) The most popular accounts of personal identity over time are of this sort. And it's clear that the personal-identity assumption implicit in the central dogma of transhumanism presupposes a psychological-continuity view: the reason why transhumanists think you could become a computer person is that they think a computer person could be psychologically continuous with you.

The most commonly proposed solution to the branching problem is to deny that someone's being psychologically continuous with you in the future suffices for you to survive. What suffices is, rather, "non-branching" psychological continuity. A later person is you just if she is psychologically continuous with you *and* there is no branching (e.g. Shoemaker 1984: 85; Parfit 1984: 207). The implication in the "uploading" case would be that as long as the psychological information from your brain is uploaded only once, the resulting person is you; but if it were simultaneously uploaded

more than once, none of the resulting people would be you. Each would be a newly created person mistakenly convinced that she was you and with false memories of your life, including the belief that she had been alive for many years. It is metaphysically possible for a person to move to a computer by "single upload" but not by "double upload."

The obvious and well-known objection to this is that non-branching requirements are arbitrary and unprincipled. The claim that you could survive single but not double uploading is surprising. And the proposal does nothing to explain *why* the occurrence of a second uploading procedure would prevent the first such procedure from moving you to a computer. Why should an event that would normally suffice to preserve your existence destroy you if accompanied by another instance of the same procedure–something that has no causal effect on the first event? What is it about the second upload that destroys you? The only answer seems to be that surviving a double upload would lead to a logical contradiction: to one thing's being numerically identical to two things. But that can't be the whole story. It cannot be merely the laws of logic that prevent us from surviving double uploading.

The current proposal faces a particularly awkward version of the branching problem. In the usual uploading stories, the brain is conveniently erased in the scanning process. But this need not be so: the relevant information could be "read off" without doing you any damage, then copied to a computer and used to create a person there exactly as before. For you it might be like having an MRI scan. Tranhumanists call this "nondestructive uploading." The result would be two people–a human person and a computer person–each psychologically continuous with you. But according to the non-branching proposal, neither would be you, as this would be a case in which two people come to be simultaneously psychologically continuous with you. And there is no other being after the transfer that you could be. It follows that you would cease to exist: nondestructive uploading would be fatal.

If this isn't already troubling enough, it raises an awkward epistemic problem. For all I know, the Martians (who have all the advanced technology that we lack) could be scanning my brain right now and copying the information to a computer, thereby creating a person psychologically continuous with me. It follows from the non-branching requirement that I could cease to exist at any moment, mid-sentence, without the slightest disruption of my mental life or physical functioning, and be instantly replaced by a new person with false memories of my life. No one would be any the wiser. It is hard to take this seriously.

Transhumanists are likely to respond by saying that it *is* possible to survive branching in this case: if the uploading procedure leaves your brain intact, you continue existing as you are, and the computer person thereby created is someone new. That, of course, sounds right. But this new proposal adds a second arbitrary and unprincipled feature to the first one. Why could someone survive "asymmetric" but not "symmetric" branching? Why, in other words, would transferring the information from your brain to a computer be "person-preserving" (as psychological-continuity theorists like to say) if, but only if, that information is gathered in a destructive way? And why, after the uploading, would you be the person with your body and not the person in the computer? The obvious answer is that you would survive as the person with your body because he or she would be materially or biologically continuous with you, and the person in the computer would not be. But the possibility of surviving ordinary, "single" uploading would imply that we can survive without material or biological continuity. Why is material continuity suddenly relevant here? The only answer would seem to be that appealing to it can avoid implausible consequences. But again, what enables me to survive asymmetric but not symmetric branching cannot be the fact that it would be implausible to suppose otherwise.

4. The Duplication Problem

Here is a second and less familiar worry about the personal-identity assumption. There has to be a difference between me and someone psychologically just like me. Someone could be a perfect psychological duplicate of me as I am at some particular time–now, say–without being me. There is a difference between a particular person and a copy or replica of that person, no matter how exact, just as there is a difference between the original Rosetta stone and a replica of it created today, no matter how exact. I don't mean a qualitative difference. A replica of the Rosetta stone might be completely indistinguishable from the original, right down to its finest atomic structure. Still, the replica would be one thing and the original would be another. The original would have been made by hand in the second century BC; the replica would have been made only today by the Martians.

So there could be a replica of Wittgenstein as he was at any moment during his life. It might resemble Wittgenstein in all intrinsic respects–a flesh-and-blood being, atom-for-atom identical to him–or it may be merely a psychological replica, with all his intrinsic mental properties but physically different. The AI assumption implies that we could create such a replica simply by programming the right sort of computer in the right way, if only we had in our possession the psychological information realized

in Wittgenstein's brain at the appropriate time. And the personal-identity assumption implies that this knowledge would enable us to upload Wittgenstein himself into a computer, abruptly resurrecting him from his quiet grave in Cambridge.

Imagine, then, that the British Wittgenstein Society somehow get access to a detailed scan of Wittgenstein's brain made shortly before his death. They propose to use it to create a psychological replica of him as he was then, so that they can put to him all the questions about his work that have accumulated in the intervening decades. (They have a long list.) A psychological replica of the man would be just as willing and able to do this as Wittgenstein himself would be. But they want a replica and not the original because they fear the interrogation will be traumatic, and they feel that Wittgenstein has suffered enough for philosophy already. The Austrian Wittgenstein Society, however, have no such scruples. They have their own copy of the scan, and want to use it to bring the great man himself back to life in order to attract foreign visitors.

If the central dogma is true, both projects are possible. The question is, what would the two societies have to do differently so that the Austrians got the original Wittgenstein and the British got a replica? It looks as if there is nothing they *could* do differently. To create a psychological replica of Wittgenstein as he was at the time of the scan, the British would have to copy the psychological information from the scan to a computer in such a way as to create a conscious, intelligent person with just the intrinsic mental properties that Wittgenstein had at a certain time in 1951. The Austrians would of course do precisely the same thing. And according to the personal-identity assumption, that would suffice to upload Wittgenstein himself into the computer. It would follow that there was no difference between bringing Wittgenstein himself back to life and creating a brand-new replica of him. Likewise, there would be no difference between your being uploaded into a computer and someone else's being newly created there. This conflicts not only with the indisputable fact that there *is* a difference between an original object and a copy, but also with the central dogma, which says that you yourself, and not merely a copy of you, could exist in a computer. Call it the *duplication problem*.

5. Why the Problems are Superficial

The branching and duplication problems are serious, and transhumanists have had little to say about them. But I don't think the problems go very deep. If uploading really is metaphysically impossible, it cannot be for these reasons–because it has absurd consequences about personal identity

over time and about the difference between originals and duplicates. These consequences are symptoms of a deeper, underlying problem.

We can see that the branching and duplication problems do not strike at the heart of the central dogma by noting that they apply equally to claims that do not involve uploading. One is that a person could travel by *Star Trek* teleportation. Suppose the teleporter works like this. When the Captain has had enough adventures on the alien planet, the teleporter "scans" him, thereby dispersing his atoms. The information gathered in the scan is then sent to the ship, where it is used to assemble new atoms precisely as the Captain's were arranged when he said, "Beam me up!" The result is someone both physically and mentally just like the Captain. And it's part of the story that the man who materializes on board the ship *is* the Captain.

If the man appearing on the ship really could be the Captain, the branching problem would apply just as it does in the case of uploading: the teleporter could produce two beings like the Captain instead of one. And if the man who appears in single teleportation would be the Captain, both men who appeared in double teleportation would be, with the impossible result that one thing is numerically identical to two things. Avoiding this problem by introducing a non-branching clause would imply that if I were scanned in a way that did not disperse my atoms and the information thereby gathered were used to assemble an exact duplicate, that would be the end of me, as it would be a case of branching.

Likewise, the information gathered in the scan could be used either to create a replica of the Captain or to recreate the Captain himself; yet the procedure for doing both these things would be exactly the same. It would seem to follow that there was no difference between a person and a replica of that person.

Another view with similar implications is Shoemaker's claim that a person could move from one organism to another by what he calls "brain-state transfer."[2] He imagines a machine that scans your brain just as in the uploading story, thereby recording all the relevant information realized in it and erasing its contents in the process. This information is then transferred not to a computer, but to another human organism with a "blank" brain, resulting in someone psychologically just like you (or as much like you as the new organism's physical properties allow). Shoemaker claims that because this being would be psychologically continuous with you, he or she would *be* you–as long as the machine copies your brain states only once and your original brain is erased. It's easy to see that the same worries about branching and duplication apply here as well.

---

[2] Shoemaker 1984: 108-111. I don't know whether any other philosopher has ever shared this view.

These views have nothing to do with uploading. They could be true even if the central dogma were false and uploading were impossible. Whatever makes teleportation and brain-state transfer impossible, if indeed they are, must be something independent of the AI and personal-identity assumptions.

Not only are the branching and duplication problems not peculiar to uploading, but there may be species of uploading that avoid the problems. Suppose the uploading process took place bit by bit rather than all at once. A small portion of your brain is scanned, and its functions, or at any rate those that are relevant to your mental properties, are duplicated in a computer. (If a computer can duplicate the functions of your entire brain, it can duplicate the functions of part of it.) The neurons communicating with the scanned brain part are then connected to the computer by radio links, and the scanned brain part itself is destroyed or disabled. The result is that your mental activity becomes scattered across parts of your brain and parts of the computer. (I don't know whether this is possible, or even whether it makes any sense; but it should be possible if the original uploading story is possible.) The procedure is then repeated with other parts of your brain one by one until all your mental activity (or all the mental activity that used to be yours) is going on in the computer and none is going on in your brain.

If the central dogma is true, it would presumably be possible to move a person from a human organism to a computer by means of such gradual uploading. If you could upload a person all at once, then you could upload a person gradually. But it doesn't look possible to construct a troubling duplication case involving gradual uploading–a case where there is no difference between moving you to a computer and merely creating a psychological replica of you there. And it would be quite a lot more difficult to construct a branching case, where there are two people, either of whom the friends of uploading would say was the original person were it not for the existence of the other.

Not that transhumanists will see this as good news. I doubt whether anyone thinks that gradual uploading is metaphysically possible but all-at-once uploading is not. There would have to be an explanation for this fact, beyond merely saying that all-at-once but not gradual uploading is subject to branching and duplication objections. It's hard to see what the explanation could be. In any event, it's clear that the metaphysical problems for the central dogma go deeper than the branching and duplication problems.

6. Material Continuity

I have said that the branching and duplication problems are symptoms of a deeper problem. What might this deeper problem be? If uploading really is metaphysically impossible, *why* is it impossible?

I think the answer is that you and I are material things: objects made up entirely of matter. That's certainly how it appears. That's why we're able to see and touch ourselves and other people. If we were immaterial, we should be invisible and intangible, which is very much *not* how it appears.

So we are material things. And a material thing cannot continue existing without some sort of material continuity. It must always be made up of some of the same matter–composed of some of the same material parts–that made it up at earlier times. A material thing can change all its parts: it can be made up of entirely different matter at different times. Owing to metabolic turnover, few atoms remain parts of a human being for long. But it cannot change all its parts at once. It cannot survive complete material *dis*continuity. It follows that you cannot move a material thing from one place to another merely by transferring information. You can't send a stone, or a shoe, or a dog as a message by telegraph (despite the joke in *Alice in Wonderland*). To move a material thing, you have to move matter–specifically, some of the matter making up that thing.[3]

But there is no material continuity in uploading. The person in the computer has none of the material parts of the human person. (Not in the usual "all-at-once" uploading, anyway.) The central dogma of transhumanism implies that you could send a person by telegraph–or, for that matter, written down in a letter. If I am right in saying that material things require material continuity to persist, then the central dogma is incompatible with our being material things.

We can make this more vivid by thinking about what sort of material things we might be. We appear to be animals: biological organisms. If you examine yourself in a mirror, you see an animal. The animal appears to be the same size as you–no bigger and no smaller. Like animals, we seem to extend just as far as the surface of our skin. Each of us seems to have the physical and biological properties of an animal: its mass, temperature,

---

[3] I believe that the material-continuity requirement derives from the further principle that material things must persist by virtue of "immanent" causation (Olson 2010). That is, they have to cause themselves to continue existing. Sometimes they need outside help–food, oxygen, medical care, that sort of thing–but the outside help can't do all the work. Corabi and Schneider (2012) argue that we cannot be uploaded because this would involve a gap in our existence. They say that material things cannot have such gaps, though I am unable to understand their argument for this claim. I suspect that if it is impossible, it's because it is ruled out by the immanent-causation requirement.

chemistry, anatomy, and so on. Nor is there any difference in behavior between a human animal and a human person. The appearance is that we *are* the animals in the mirror.

Our being animals is clearly incompatible with the central dogma. You cannot move a biological organism from a human body to a computer by scanning its brain and "uploading" the information thereby gathered. Scanning may leave the organism unharmed. Or it may damage it, perhaps even fatally. It may even completely destroy the organism by dispersing its atoms (as the *Star Trek* transporter does). But no matter what form the scan takes, the organism stays behind. It may remain unchanged, or be damaged or killed or completely destroyed, but it is not converted into information and transferred to the computer. You couldn't point to an electronic computer and say, "That thing was once a microscopic embryo composed of a few dozen cells."

So if you and I are organisms, it would be metaphysically impossible to upload us into a computer. Of course, we might be material things other than organisms. A few philosophers say that we are brains, or parts of brains. (Parfit 2012; see also Olson 2007: 76-98) Each of us is literally made up entirely of soft, yellowish-pink tissue and located within the skull. But it is no more possible to upload a brain into a computer than an animal. The scanning does not remove the brain from the head and convert it into information. The brain is a physical object, like a heart or a kidney. It may remain unchanged in the scanning process, or it may be damaged, or even completely destroyed by having its atoms dispersed, but it is not converted into information and transferred to the computer. You couldn't point to an electronic computer and say; "That thing was once a three-pound mass of soft tissue."

If the central dogma is true, then, it follows that we can be neither organisms nor brains. Not only could we not be organisms or brains once we have been uploaded, but we could not be organisms or brains even now. And not only are we not organisms or brains *essentially*. We are not organisms or brains even accidentally or contingently. The central dogma implies that a human person has a property that no organism or brain has, namely being uploadable into a computer by a mere transfer of information.

Suppose this attractive account of our metaphysical nature were true: we are biological organisms, or perhaps brains. What's more, all conscious beings, at the present time anyway, are organisms or brains. That would explain why a human person cannot be uploaded into a computer, contrary to the personal-identity assumption: because we are organisms or brains, and it is metaphysically impossible to move any material thing to a computer simply by transferring information.

7. The Pattern View

I have argued that we are material things, and that material things cannot persist without material continuity. As there is no material continuity in uploading, that explains why uploading is metaphysically impossible. (The same goes for *Star Trek* teleportation and Shoemaker's brain-state transfer.) There are two ways of defending the central dogma against this argument: to deny that we are material things, or to deny that material things require material continuity to persist. (I don't suppose anyone will argue that uploading is possible only in gradual cases where there *is* material continuity.) I will consider these proposals in turn.

To deny that we are material things is to deny that we are made up entirely of matter. And that is to say that we are partly or wholly made up of something else: we are (at least partly) *im*material things. And this is something that transhumanists often do say. Specifically, they often say that a person is a sort of *pattern*. Bostrom claims that in the future it will be possible for us to "live as information patterns on vast super-fast computer networks" (2016). Ray Kurzweil says that owing to the fact that living organisms constantly exchange matter with their surroundings:

> all that persists is the pattern of organization of that stuff..., like the pattern that water makes in a stream as it rushes past the rocks in its path....Perhaps, therefore, we should say that I am a pattern of matter and energy that persists over time. (Kurzweil 2006: 383)

And Daniel Dennett suggests that "what you are is that organization of information that has structured your body's control system." (1991: 430; see also 1978) Perhaps the very same pattern or form of organization could be instantiated or realized first in a biological organism and then in an electronic computer. And if this is possible, the scanning-and-uploading process that transhumanists imagine would be the way to do it. This may not be true of all patterns instantiated in the brain: those involving fluid dynamics or ion transfer across membranes are probably not transferable to an electronic substrate. But perhaps those patterns relevant to psychology are.

The proposal has to be that a person–the author of this paper, for instance–is literally a pattern of some sort. It is not merely that to be a person is to instantiate a certain sort of pattern, or that for a person to exist is for such pattern to be instantiated. These may or may not be sensible claims, but they do nothing to explain how a person could move from a human organism to an electronic computer. (They are compatible with our being organisms.) The suggestion is that we are not things that instantiate certain patterns, but that we are those patterns ourselves.

Could a conscious, thinking being be a pattern? The question is hard to think about because the word "pattern" is so nebulous. (I suspect that this lack of clarity is what has encouraged transhumanists to speak casually of our being patterns.) What sort of thing is a pattern? The assumption has to be that it is not a material thing of any sort, but rather something that can move from one material thing to another by a transfer of information. But that doesn't tell us much.

As far as I can see, a pattern would be a sort of property or relation: a universal. Different concrete objects, or collections of objects, can exemplify the same pattern, just as different flowers can have the same colour. All the copies of *Moby-Dick* in the original English have the same pattern of words and letters. (Or nearly the same. Let us ignore irregularities in typesetting, different locations of line and page breaks in different editions, and the like.)

The view that we are patterns, so construed, would solve the branching and duplication problems. If you were uploaded twice over, the result would be not two people, but only one: the same pattern would be present in two different computers. (Olson 2007: 146f.) There would be two *instances* of the pattern–that is, two physical things patterned in the same way–but there would be only one pattern in both. They would be the same person in the way that two physical volumes might be the same book–*Moby-Dick*, say. So double uploading would not have the impossible result that one thing is numerically identical with two things.

The proposal would solve the duplication problem by implying that a copy or replica of a person, if it instantiates the relevant pattern, *is* that person and not a replica. Both the Wittgenstein created by the British and the Wittgenstein created by the Austrians would be the original Wittgenstein born in 1889. Or more precisely, both physical objects would be instances of the same person: the solution assumes that neither physical object would itself be a person.

But the pattern view is impossible to take seriously. Suppose we ask which pattern a given person might be. If there are such things as patterns, this human organism now instantiates many of them. There is, for instance, the pattern consisting of the current orientation of my limbs, and the pattern formed by the flow of material through my gut. Which pattern am I? Since I am conscious and thinking, I must be the one that instantiates those mental properties. The pattern view presupposes that of all the patterns instantiated here, one of them, and only one, can think. That's because there is just one thinking being here, namely me. But which of those patterns is the one that thinks? Of all the patterns the organism instantiates, what could make just one of them conscious? I have no idea how

to answer this question. It's no good saying that to be conscious or intelligent *is* to instantiate a certain pattern. Although that may be true, it would imply that the organism was conscious, since it is the thing instantiating the pattern. That would make typical human organisms conscious and intelligent, yet not uploadable–precisely what the pattern view was meant to avoid. The proposal has to be that no material thing could possibly have any mental property.

But perhaps the most obvious problem for the pattern view is that universals don't *do* anything. They don't change. And this prevents them from thinking or being conscious. When we speak of changing the pattern or arrangement of chairs in the room–from square to circular, say–we mean rearranging the chairs so that they instantiate a *different* pattern from their current one. A single pattern cannot be first square and then circular. It can change only in the way that the number seventeen changes by ceasing to be the number of chairs in the room when we move one next door: mere "Cambridge change," as they say. A universal cannot undergo any real, intrinsic change.

But if I know anything, I know that *I* undergo real change. I am sometimes awake, for instance, and sometimes asleep. That I change intrinsically follows from the fact that I am conscious and thinking. No person–even a computer person–could be a pattern. A thing that changes can at best be a particular *instance* of a pattern and not the pattern itself: a concrete thing that is patterned or organized or arranged in that way. The claim that a person is an instance of a pattern is entirely harmless. Every concrete object is an instance of some pattern or other (still supposing that there are such things as patterns). But again, the claim that we are instances of patterns tells us nothing about how we could be uploaded into a computer.[4]

8. The Constitution View

Turn now to the proposal that we can survive complete material discontinuity despite being entirely material things. One view of this sort incorporates the thought that a human person is not an organism, but rather a material thing "constituted by" an organism. Each of us stands to an organism in the way that a clay statue stands to the lump of clay making it up. A human person is made of the same matter as the organism we might call its body, and physically indistinguishable from it. But the person differs from the organism in its modal properties: the person, but not the organism, persists by virtue of psychological continuity. So in Shoemaker's brain-state transfer story, a person would be constituted first by one organ-

---

[4] For more on the pattern view, see Olson 2007: 145-149.

ism and then by another, much as a statue that got an arm replaced would be constituted first by one lump of clay and then by another. And perhaps in uploading, a person could cease to be constituted by any organism, and come instead to be constituted by some part of a computer.[5] This need not imply that *all* material things can survive without material continuity. It might be impossible for an organism or a lump of clay. But material things of our sort can.

There is a large and ongoing debate over the merits of the constitution view, independent of whether it would allow uploading.[6] But the view is unlikely to appeal to transhumanists. For one thing, it does nothing to explain *how* it is possible for a material thing to survive without material continuity. If it seems absurd to suppose that a thing made entirely of matter could be sent as a message by telegraph or dictated over the phone, the proposal tells us nothing about why this appearance is misleading. It says, of course, that personal identity over time consists in some sort of psychological continuity, generously construed so that it does not require material continuity. But this simply asserts that material continuity is unnecessary, and does nothing to address the strong conviction to the contrary. What's more, the claim is entirely independent of the constitution view. If it's a sensible thing to say, it's sensible whether or not we are constituted by organisms.

Nor does the proposal suggest any solution to the branching and duplication problems. If uploading could bring it about that I ceased to be constituted by an animal and became constituted instead by a computer, then it could apparently bring it about that I became constituted simultaneously by one computer and also by another, making me numerically distinct from myself. And there would appear to be no difference between a computer's constituting me as a result of uploading, on the one hand, and a computer's constituting someone else just like me, on the other, and thus no difference between a person and a mere copy of that person.

9. The Temporal-Parts View

The best way of defending the central dogma may be to appeal to the ontology of temporal parts.[7] It consists of two principles. First, all persisting things are composed of arbitrary temporal parts. A temporal part of some-

---

[5] Both Baker and Shoemaker believe that we are constituted by organisms and that we can survive without material continuity (Baker 2005; Shoemaker 1984: 108-114, 1999). Given the AI assumption (which Baker accepts; cf. 2000: 109), it follows that I could become constituted by a computer through uploading.
[6] For a summary, with references, see Olson 2007: 48-75.
[7] This is a difficult topic. I discuss it at greater length in Olson 2007: 99-128.

thing is a part of it that takes up "all of that thing" at every time when the part exists. Barry Manilow's nose is a part of him, but not a temporal part, because it doesn't take up all of him while it exists. His adolescence or his first half, though, if there are such things, would be temporal parts of him. A temporal part of something is exactly like that thing at all times when the part exists. It differs from the whole only by having a shorter temporal extent. To say that persisting things are composed of *arbitrary* temporal parts is to say that for any period of time when a thing exists, there is a temporal part of it existing only then.

The second principle is unrestricted composition: for any entities whatever, there is a larger thing composed of them. (Some things, the xs, compose something $y =_{df}$ each of the xs is a part of y, no two of the xs share a part, and every part of y shares a part with one or more of the xs.) So if there are such things as Barry Manilow's nose, Plato's fourth year, and Yugoslavia, then there is also an object scattered across space and time that is made up of those three things. Both principles are, of course, highly controversial. Together they imply that every matter-filled region of spacetime is exactly occupied by a material thing. This is what Quine meant when he said that a physical object "comprises simply the content, however heterogeneous, of some portion of space-time, however disconnected and gerrymandered." (1960: 171)

It follows from the principle of arbitrary temporal parts that I have a temporal part extending from the beginning of my existence until midnight tonight, and that my computer has a temporal part extending from that time until the computer's demise. And it follows from unrestricted composition that there is something composed of these two objects: a material thing, given that both I and my computer are material things. It is conscious and intelligent until midnight tonight, when it "jumps" discontinuously from me to the computer. From then on it is not conscious or intelligent. (Splendid though my computer is, its powers are limited.) If my computer really did have the right mental capacities, though, then the being jumping from me to it would remain conscious and intelligent. In fact such a being would make this jump at every moment at which both the computer and I are conscious, with or without any sort of "uploading"– that is, any transfer of information from the organism to the computer. That's because the computer and I are each composed of arbitrary temporal parts, and any two of them compose something. Any pair consisting of one of my temporal parts and one of my computer's, provided they don't exist simultaneously, will jump from one of us to the other.

So according to the ontology of temporal parts, it is perfectly possible for a material thing–even a conscious, intelligent one–to persist without

material continuity. It does not follow from this, however, that a *person* could move from a human body to a computer. To secure this claim–the personal-identity assumption–such beings would have to count as people. And on the temporal-parts ontology, having the mental capacities characteristic of personhood–intelligence, self-consciousness, and the like–does not suffice for being a person. Many of my temporal parts, such as the one that extends from midnight last night till midnight tonight, have those mental capacities but are not people. No person now writing these words is going to perish at the stroke of midnight, without any injury or other disruption of his mental or physical activities. At any rate, few temporal-parts theorists think so. (Sider [1996] is an exception.) Not just any rational and self-conscious being is a person.

We can see this point by noting that the ontology of temporal parts entails the existence of a thing composed of the temporal part of me extending from my beginning till midnight tonight and the temporal part of you extending from that time till your demise: a conscious, intelligent being jumping from me to you. But this being is not a person, and its existence is of no practical or metaphysical interest. If I knew that I was going to be shot at dawn, the conviction that this being was going to survive that event would be no more comfort me than the thought that you were going to survive it.

So the temporal-parts ontology implies that conscious, intelligent beings could move from human bodies to computers by uploading. There is no metaphysical mystery about this–or at least none beyond that inherent in the temporal-parts ontology itself and the AI assumption.

The proposal would also solve the branching and duplication problems. Suppose my brain is scanned (and thereby erased) and the information gathered is uploaded simultaneously into two computers. Two people emerge from the process. Both, temporal-parts theorists can say, would be me. How could two things be one thing? The reply is that in this case there are two people all along, who share their pre-upload stages but not their post-upload stages. (Call the short-lived temporal parts of people "person stages.") These people begin to exist when I do and share all the events of my life until the uploading takes place. During that period there is no difference between them. But afterwards they live in different computers and lead independent lives. This is a consequence of the claim that there is a being composed of my pre-upload stages and the post-upload stages of the one computer, and also a being composed of my pre-upload stages and the post-upload stages of the other computer, together with the assumption that such stages are connected in the way that makes for personal identity over time–that is, that makes them compose a person. The two people are

like railway lines that share their tracks for part of their length and diverge elsewhere.

What about the duplication problem? What would be the difference between bringing Wittgenstein himself back to life, by programming a computer with the psychological information from his brain, and creating a psychological replica of him by that means? According to the temporal-parts ontology, there is no deep metaphysical difference between originals and replicas. Suppose we somehow produced a computer person psychologically identical to Wittgenstein as he was shortly before his death. The temporal-parts ontology would imply that there are two conscious, intelligent beings in the computer, insofar as the intelligent computer stages are parts of two such beings. One was born in 1889 and wrote the *Tractatus Logico-Philosophicus*. The other began to exist only just now. They share their current stage, but the 1889-to-1951 stages are parts of the first and not of the second. If the first being counts as a person, then we have resurrected Wittgenstein himself. If the second is a person, then we have merely created a replica of him. (Requiring a person to be a *maximal* aggregate of appropriately interconnected stages—a thing, each of whose stages is appropriately connected to every other, but which is not a part of any larger such thing—would rule out their both being people.) But which of these is the case is not a metaphysical question, but simply a matter of how we use the term "person."

According to the temporal-parts ontology, then, conscious, intelligent beings could move from human bodies to computers via uploading. And these beings would be people, vindicating the personal-identity assumption, just if the stages of those beings would relate in the way that would amount to their composing a person. What relation is this? Transhumanists will say that it is some sort of psychological continuity or connectedness. Perhaps a person is a maximal aggregate of psychologically interconnected person stages: that is, a being composed entirely of person stages, each of whose stages is psychologically connected to every other, and which is not a part of any larger such being. (Lewis 1976) And we might say that two person stages are psychologically connected just if the mental properties of one of them depend causally in the right way on those of the other.

It's clear that the post-upload stages of a computer person could have mental properties that depend causally on those of the pre-upload stages of human people. But would they depend in the right way–the one that would make the beings who move from human being to computer count as people? That looks doubtful. An attractive thought is that stages are parts of the same person only if they are connected by relations of practical concern: if one has "what matters" to the other. (Parfit [1984: 262] calls this

connection "relation *R*".) In other words, a future person is me only if I now have a reason to care about his or her welfare then–a reason I should have even if I were completely selfish and would not lift a finger to save my own mother from unbearable pain. This may also imply that that future person would be morally responsible, then, for the things I do now (in the absence of the usual excuses, such as insanity), that he or she would then deserve compensation for my efforts now, and so on.

Transhumanists are likely to accept this requirement. Their claim that we could become computer people is not meant to be of merely theoretical interest. They think it would matter to us, practically speaking, if we were uploaded into computers: it could benefit us, and we have a reason to try to bring it about. If we could become computer people but their welfare would be of no practical importance to us, then we should have no selfish reason to upload ourselves, no matter how wonderful the life of a computer person would be. Uploading ourselves would be no better, for us, than creating psychological duplicates of ourselves in computers.

In fact, transhumanists would rather say that computer people could have what matters practically to us without being us than that computer people could be us without having what matters. That is, if the personal-identity assumption turned out to be false and it was metaphysically impossible for us to become computer people, they would retreat to the claim that computer people could at least have what matters: they could bear to us those relations of practical concern that give us a reason to care about our own future welfare. Even if we cannot literally be uploaded, they will say, it could be that as far as everything we care about is concerned, it's as good as if we could. My worries about the personal-identity assumption are of interest only to metaphysicians. The rest of us can ignore them.

So according to the temporal-parts ontology, we could be uploaded into computers only if beings that move from human bodies to computers via uploading could count as people. And they could count as people only if each of their stages has the mattering relation to every other, or at any rate only if their post-upload stages have what matters practically to their pre-upload stages. Is this the case? It doesn't seem so. Consider once again the case of nondestructive uploading. Suppose I am kidnapped by bad people, who are going to scan my brain and upload the information, resulting in both a computer person psychologically just like me and a human person entirely like me (and materially continuous with me to boot). Then they are going to torture one of these people. The magnitude of the suffering will be the same in either case. But for some reason they allow me to choose which person is tortured: the human person with my body or the computer person. (Suppose I accept the AI assumption: I don't doubt that they could create and torture a computer person.)

If uploading preserves what matters, I ought to be indifferent. But I should be anything but indifferent. Even if I were completely selfish, I would far rather that the computer person be tortured than the human person. I suspect, in fact, that if I were completely selfish I should be indifferent about the welfare of the computer person. My only concern would be the welfare of the human person. And I doubt whether these attitudes are peculiar to me.

Or imagine that the bad people work out how to scan people's brains remotely without their noticing. They then upload the information from the scan into a computer, creating someone psychologically identical to the unsuspecting victim as she was when she was scanned. This being is then tortured. Suppose the bad people have been active in my neighbourhood, and I believe there is a real chance that they will scan my brain tonight and torture the resulting person. If uploading preserves what matters, I ought to be just as worried about this as I should be if I thought there was a real chance that the human person who will wake up in my bed tomorrow will be tortured. But I should find the second case far more worrying.

Someone might suggest that *destructive* uploading preserves what matters practically, even though nondestructive uploading does not. A computer person produced by scanning and uploading the information in my brain would have what matters to me if, but only if, there is not also a human person then who is both psychologically and materially continuous with me. And this is not because the computer person *is* me just if no such human person issues from the procedure: we are assuming that a person moves from my body to a computer in either case. The suggestion is that the computer person would be me (by sharing my current stage), but whether I should have any selfish reason to care about his welfare depends on what *other* people existing after the procedure would also be me (by sharing my current stage). But no philosopher I know of has ever held this view. Parfit's famous arguments in Part 3 of *Reasons and Persons* presuppose that what matters cannot depend on what we might call extrinsic factors, and none of his many critics have questioned this assumption.

It appears, then, that uploading does not preserve what matters practically. Assuming that stages are parts of the same person only if they are connected by relations of practical concern, it does not look as if a person could move from a human organism to a computer by uploading. The procedure may move *some* material thing from a human being to a computer–this is guaranteed by the ontology of temporal parts–but not a person. The personal-identity assumption looks false even given the temporal-parts ontology.

So it looks rather doubtful whether the temporal-parts ontology can save the central dogma of transhumanism. For the same reason, it looks doubtful whether computer people could have what matters to us in identity: whether having psychological duplicates in computers would be just as good for us, practically speaking, as literally moving there ourselves. But I'm not very confident about this. It may be that I am simply wrong about what matters practically–about what would be in my own interest–and that my reactions to the imagined cases are mistaken. In any event, transhumanists have work to do.

REFERENCES

Baker, L. R. (2000). *Persons and Bodies: A Constitution View*. Cambridge University Press.

Baker, L. R. (2005). "Death and the afterlife." In W. Wainwright (ed.) *Oxford Handbook for the Philosophy of Religion*. Oxford University Press: 366-39.

Bostrom, N. (2001). "What is transhumanism?" <http://www.nickbostrom.com/old/transhumanism.html> accessed 9.8.16.

Bostrom, N. *et al.* (2016). Transhumanist FAQ, 3.0. <http://humanityplus.org/philosophy/transhumanist-faq/>, accessed 9.8.16.

Chalmers, D. (2010). "The singularity: A philosophical analysis." *Journal of Consciousness Studies* 17 (9-10): 7-65.

Corabi, J. and Schneider, S. (2012). "The metaphysics of uploading." *Journal of Consciousness Studies* 19 (7-8): 26-ff.

Dennett, D. (1978). "Where am I?" *In Brainstorms*. MIT Press.

Dennett, D. (1991). *Consciousness Explained*. Little, Brown.

Kurzweil, R. (2006). *The Singularity is Near: When Humans Transcend Biology*. Duckworth.

Lewis, D. (1976). *Survival and identity*. In A. Rorty (ed.) *The Identities of Persons*. University of California Press. (Reprinted in his *Philosophical Papers* Vol. I, Oxford University Press, 1983.)

Olson, E. (2007). *What Are We?* Oxford University Press.

Olson, E. (2010). "Immanent causation and life after death." In G. Gasser, (ed.) *Personal Identity and Resurrection*. Ashgate: 51-66.

Olson, E. (2015). "Life after death and the devastation of the grave." In M. Martin and K. Augustine, (eds.) *The Myth of an Afterlife*. Rowman & Littlefield.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Parfit, D. (2012). "We are not human beings." *Philosophy* 87: 5–28.

Quine, W. V. O. (1960). *Word and Object*. MIT Press.

Quine, W. V. O. (1999). "Self, body, and coincidence." *Proceedings of the Aristotelian Society, Supplementary Volume* 73: 287-306.

Shoemaker, S. (1984). *Personal identity: A materialist's account*. In S. Shoemaker and R. Swinburne *Personal Identity*. Blackwell.

Sider, T. (1996). "All the world's a stage." *Australasian Journal of Philosophy* 74: 433-53.

van Inwagen, P. (1978). "The possibility of resurrection." *International Journal for Philosophy of Religion* 9: 114-121.

2. Embodied And Extended Self

MILJANA MILOJEVIĆ

1. Introduction[1]

Although the notion of self can be conceptualized in many different ways, the focus of this paper will be on one particular understanding of self, namely the one that closely connects it to the concept of personal identity. On this view self is what makes us, me or you, different from others. It is the locus of our thinking, perceiving and acting, a subject that persists through time as a continuous entity having a unique identity. Such a notion of self is to be distinguished from the notion of a person, although the two are tightly related. When we are searching for the criteria for personhood we are looking for general traits of a certain kind in virtue of which entities of this kind can be considered as being subjects of rights and obligations. On the other hand, establishing criteria of personal identity is identifying properties that are responsible for a being's awareness, continuity, and uniqueness. It could be said that having or being a self is a prerequisite for being a person, so insights about the nature of personhood and self will often overlap.

There are many interconnected questions that every theory of self or personal identity tries to answer: what is the ontological nature of the self; what are the necessary conditions for personal identity; what is the relation between the self and the body; what are the persistence criteria of the self; etc. Olson (2009) distinguishes three groups of these questions, which are, in order, concerned with: persistence, population, and personal ontology. The first group is focused on issues relating to diachronic identity, and questions about biological death and persistence of self, among others. The second is related to concerns about the extension of the term "self": Is it determined by biological or psychological criteria? Can there be more than one person in one body? and the like. While the third group of questions

centers on metaphysical topics: Are selves independent substances and of what kind? Are they states or properties of something else? Are they just functional properties? etc. All these questions are interrelated and particular answers to each of them will have consequences for the answers to others. The question of immediate interest to us, and the topic of this paper in general, will be the role of the body and the environment in determining personal identity, or in the constitution of the self. We will argue that the material body of the subject as well as some parts of his environment play a much greater role in the constitution of the self than has traditionally been thought. In order to defend this claim we will endorse a psychological view of the self and a functionalist outlook on mental and cognitive states as well as on personal identity. As a consequence of employing a special kind of functionalism which allows for extended or widely realized mental states, we will change the traditional extension of the term "self," and incorporate environmental and bodily factors in the determination of personal identity and the persistence of the self. We will start with the ontological status of the self, which will have direct consequences for answering the population and persistence questions with respect to personal identity.

2. The Ontology of the Self and its Relation to the Body

One of the most prominent divisions of ontological views of the self is the one which contrasts immaterialism[2] and animalism.[3] These views are usually seen to be committed to the assumption that selves are certain types of things or substances, so we can call such views substantive or essentialist views of the self. Nevertheless, they can also be conceptualized as different answers to the question: What is the relation of the self and the body? (Cassam 2011) This question is of immediate interest to us, as the plausibility of the hypothesis that the self is embodied or extended will depend on the kind of answer given when asked: Is the self detachable and separate from the body, partially or wholly constituted by it, or identical with it?

According to immaterialism, the self or the soul is something different from matter, and consequently different from the body that a subject has or is. Such a view most often comes in a substance dualism form, not to be confused with property dualism, or views such as Hume's, that a self is nothing more than a bundle of ideas. Plato, Descartes, and Leibniz all

---

[2] Immaterialism is to be understood in relation to the ontological status of the self, and to be distinguished from Berkeley's immaterialism, which postulates that there is no matter. Immaterialism about the self is thus compatible with both metaphysical dualism and immaterialism.

[3] A third view is sometimes added to the list, constitution view described by Baker (2000).

thought that the soul or the self is something independent of the body and robust enough to be detachable from it. Let us just briefly remember that Descartes articulated not one, but several arguments in the *Meditations* and *Discourse on Method* for the distinctness of the mind, the soul or the self, from the matter. According to him there is a real distinction between our minds and our bodies, where the immaterial self is indivisible and indubitable, unlike material things. On these views the body does not have any constitutional role in the creation or shaping of the self. A self is typically situated in a body[4], but detachable from it, and acts as a separate entity.

On the other hand, animalism, having its heritage in Aristotle's theory of the soul, claims that we are a kind of material substance; we are selves in virtue of being biological organisms: more precisely, we are selves because we are human animals. In claiming this, animalism is presupposing that selves are not detachable from the body, or different from it as separate substances.[5] Both views, immaterialism and animalism, identify selves with substances which have essential features, and both views face a multitude of problems. Immaterialism has to deal with the question of interaction between the self and the body, and to provide plausible criteria for identification of particular selves, while animalism, which equates personal identity with bodily identity, has to explain what is so special about human beings, whether selves have the same temporal existence as bodies, and by what criteria we can attribute personhood and selves to non-animals and non-humans, etc.

Starting from the question about the possibility of an embodied self, or a self that is deeply dependent on bodily features, which could at the same time be extended into the environment by having non-biological constitutive parts, it seems that none of these alternatives can be a stepping stone in conceiving these possibilities as well founded. The immaterialist view is explicit in rejecting the body as constitutive of the self, while the animalist view rejects *a priori* the possibility of having non-biological parts because it identifies us as biological organisms. So this brief look into essentialist views leads us to conclude that they cannot provide a basis for claims that

---

[4] For Leibniz matter is not a separate substance, but appears as a system of pre-established relations of perceptions of monads, which are the basic constituents of reality. In that sense, a self is not situated in the body, but it perceives itself as being situated in one.

[5] Aristotle's view (*De Anima*) is somewhat different from modern animalist views in that it postulates that the soul is the form of the body, and not the biological organism itself. Also, the soul is the principle of all living creatures: plants, animals and humans. Contemporary animalist views have a different starting point and they say that humans who are persons are animals.

the self is embodied and sometimes extended, for which we will provide reasons later. Fortunately, we do not have to accept such strong ontological commitments, and to identify selves with certain substances or natural kinds in a strong sense. Nor do we have to be as skeptical about the self as Hume, who reasoned that because we derive our idea of the self from impressions which do not themselves persist, there is no persisting self which underlies the impressions and ideas. The self could be, on the other hand, functionally identified, and ultimately realized by a variety of processes or states, or a structure with no essential features which would constitute a natural kind. Instead of looking for an object or a kind that we call "self," we can focus on functions that we expect an entity called a "self" to perform. These functions are most commonly conceived as psychological functions. A self is an entity capable of thinking, perceiving and acting, and which persists by being psychologically continuous. This is why a psychological view of the self tries to establish appropriate functions that identify an entity as being a self, where these functions could be satisfied by a specific structure, a set of processes or states, or a physical system. Although we do not want to adopt the substantivist assumption, there is still a basic insight from these considerations that selves are either materially or immaterially realized. So even if we do not have to accept that selves are what they are in terms of having specific immaterial or physical/biological properties, we can still be committed to a weaker ontological claim that selves are instantiated as physical or immaterial structures. These weaker claims do not force us to adopt the existence of immaterial souls, or the animalist view that biological creatures are identical to selves.

We should note that it would be *ad hoc* to assume that we need to adopt a psychological view of the self simply because immaterialism and animalism do not support the hypothesis that there are embodied and extended selves. There are independent reasons for adopting a psychological view of the self, and it could even be said that such a view is predominant in the contemporary debate. This view is often seen as the account which, apart from having fewer ontological commitments, solves some of the problems immaterialism and animalism face. For instance, if animalism is correct and a self is identical with a specific human animal, then there cannot be two persons in one biological body. On the other hand, split-brain cases have often been interpreted since the 1960s as instances of two persons in one body, most notably by neurobiologist Roger Sperry and cognitive neuroscientist Michael Gazzaniga, who thoroughly investigated patients with split-brains throughout the years. Namely, it was noticed that after a surgical procedure called calloscotomy, a kind of commissurotomy, introduced in the 1940s to treat epileptic seizures, two personalities were

sometimes created. (Sperry 1968, Gazzaniga 1995) In this procedure the corpus callosum is severed, leaving the right and left hemispheres of the brain without their main connection. Sperry and Gazzaniga researched the phenomena for many years and devised various experiments for testing it. In some cases calloscotomy left patients with two conflicting brain hemispheres with differential behaviors: the left hand fighting the right one in buying groceries, giving different answers about goals and desires if the question was presented to a different hemisphere (left or right eye), unable to verbally articulate the answer but drawing it by the left hand, etc. This research and its conclusions lead us to believe that bodily or brain identity are not the best guides to personal identity.

On the other hand, immaterialism identifies a person or a self with a particular soul as a substance, and so allows that as long as the soul is numerically identical the self remains unchanged. But it was already noticed by Locke that such an identity, or a bodily identity, is not sufficient for personal identity. In order to illustrate this point he provided a thought experiment in which the thoughts and memories of a prince were transferred to a cobbler. The intuition we have is that the prince is now in the cobbler's body. He could be in the cobbler's soul as well, if souls exist. The criterion of personal identity is not in the possession of the same body or the same soul, but in being psychologically continuous. Souls or immaterial selves could have all their memories and thoughts erased, yet it seems unlikely that such a soul would retain the same self.

It seems that psychological determination of the self avoids these problems. If psychological continuity is what establishes personal identity, then in the split-brain cases we could distinguish two persons in one body if they are psychologically separate, and we can also account for the potential transfer of the self from one body to another. But if selves are transferable, how can we plausibly talk about embodiment? It seems that there is a certain tension between functionalism, embodiment, and the extension of the self. This brings us to the question of multiple realizability, and its conditions and consequences for the theory of the self.

3. Psychological Continuity, Functionalism, Multiple Realizability and Embodiment

The psychological view of the self is taken to be a functionalist theory, as it is mainly focused on certain psychological functions rather than on what kind of thing, ontologically speaking, the self could be. This does not mean that psychology could not be paired with an appropriate ontology or determination of the realizers of these functions, as it is done, for example, in neuroscience. So if we are taking the psychological stance towards personal

identity we have to distinguish between psychological continuity as a criterion for personal identity functionally determined – which will include questions about the role of memory in retaining psychological continuity, or whether in cases of split-brains there are one or two persons in one body – and realizers of the functions essential for psychological continuity, which constitute the ontological basis of the self. It is also useful to distinguish between two forms in which psychological functionalism can come with respect to functional and ontological, or higher and lower level properties. Namely, role functionalism identifies mental states and processes with higher-order properties such as functional properties, while realizer functionalism identifies those same states with the typical realizers of the specified functions. In the debate about personal identity it seems that the psychological account is often taken to be a kind of role functionalism, or at least a functionalist theory with no special ontological commitments. An exception is made in the so-called hybrid brain theory, which is seen as combining both aspects of psychological and physical identity criteria in equating personal identity with brain identity. Brain theory is then, according to the distinction introduced, just one instance of realizer functionalism, which claims that appropriate realizers of psychological continuity are neural states or neural connections in the brain. We will defend a kind of realizer functionalism, but one which will broaden the neurological base of potential realizers.

Let us go back to the question of multiple realizability. If psychological continuity is taken to be functionally determined, it is easy to imagine cases in which one person is psychologically continuous with several people, which would violate the premise that personal identity is a one-one relation, or that personal identity is a type of numerical identity. Another problem is that it seems that if different bodies can instantiate the same person, then the thesis of embodiment cannot be paired with functionalism. These problems are highly interconnected and rest on the premise of multiple realizability, which both role and realizer functionalism about the mental endorse. The illustration of these problems comes in the form of body switching, branching and reduplication scenarios, but on the other hand problems will also arise if we accept a specific view on embodiment that presupposes the incarnation of the mind assumption. We will now address some objections to the psychological view: one will challenge our intuitions, the second will point out the violation of one-one nature of personal identity, and the last objection will claim that functionalism cannot accommodate the embodiment of the mental or the self.

Sydney Shoemaker (1984) describes a hypothetical situation in which two men, Brown and Robinson, undergo a kind of surgical procedure in

which, for some reason, it is needed to remove the brain of the person and later return it into the body. Unfortunately for Brown and Robinson, during the procedure their brains get mixed up, and Brown's brain is placed in Robinson's body. A similar scenario, in line with the multiple-realizability premise, is offered by Snowdon (1995) and Cassam (2011), where recorded brain states of one person are transferred to another person's body. The question is: After discovering that Brown's brain, or his thoughts and memories, are in Robinson's body, do we call this person Brown, Robinson, or perhaps Brownson? The answer that proponents of psychological view give is that it is Brown who is now in Robinson's body, while animalists would be inclined to say that something awful has happened to Robinson, and that Brown's thoughts are now imposed on him. The same ambiguity about how we understand personal identity is reflected in the question of whether to name a procedure in which the head of one person is attached to the body of another (about to be attempted on humans as a treatment for quadriplegia) "full-body transplant" or "head-transplant." While a number of authors insist that it is the mental states that are responsible for calling someone a person in the first place, a second group emphasize physical identity. These scenarios are not conclusive and can be seen as intuition pumps. What is conclusive, though, is that we do not have clear general intuitions about personal identity, or at least that there are reasons for believing that both mental states and bodily states constitute personal identity. It is worth noticing that the two interpretations are seen as portraying the conflict between the psychological view and animalism. Nevertheless, it is better to conceive these different interpretations as two possible views on the role of the body in constituting personal identity. The psychological view is taken *a priori* to be at odds with embodiment, which is yet to be determined. In other words, it is assumed that mental states can be realized in isolated brain matter or in artificial digital media, and thus detached from the body. On the other hand, if our hypothesis about the embodied and extended self turns out to be a plausible contender in the debate, it will encompass both intuitions, the intuition about the primacy of mental states in determining personal identity and the intuition about the role of the body in its constitution. The scenario with brain transplantation would be reinterpreted as a case of only partial transfer of the self, and the possibility of recording and storing mental states on a digital device would be questioned.

Another group of scenarios have greater consequences for the psychological view as the correct view of personal identity, namely the branching and reduplication scenarios we find in Parfit (1984). Both kinds of scenarios include a teletransporter that is capable of transferring people from Earth

to Mars. Just as in the *Star Trek* series, a person is teletransported from one location to another. The body of the person who is transferred is destroyed in the starting location and duplicated at the arrival location. According to the psychological continuity criterion, the person recreated at location 2, the location of arrival, is identical to the person destroyed in location 1, the point of departure. All the relevant mental states are now physically realized in the replica's body. If we were to hold a physical criterion of personal identity, the replica would have to be considered a different person from the person destroyed on Earth. So engaging in this thought experiment leads us to think that the psychological continuity view, which allows for multiple realizability, scores better with our intuitions about personal identity. But the story does not end here. Let us consider a variation of this scenario. Derek, person A, is about to be teletransported to Mars, but the teletransporter malfunctions and Derek, person B, continues to exist on Earth, while a replica of him, person C, is produced on Mars. According to the psychological continuity criterion, we would have to hold that A is identical to both B and C, violating the principle that personal identity is a one-one relation. A is now physically realized in both B and C, which is untenable and makes the psychological continuity criterion insufficient for determining personal identity. The application of the physical criterion of personal identity, in contrast, does not violate this principle, as according to it, A would be identical only with B, and not with C. Parfit has his own solution to this problem: excluding the branching cases, as just described, as the cases of personal identity. He claims that the following conditions have to be met in order to establish personal identity:

> (1) There is psychological continuity if and only if there are overlapping chains of strong connectedness. X today is one and the same person as Y at some past time if and only if (2) X is psychologically continuous with Y, (3) this continuity has the right kind of cause, and (4) it has not taken a "branching" form. (5) Personal identity over time just consists in the holding of facts like (2) to (4). (Parfit 1984: 207)

In cases where branching occurs we can only talk of survival and not of personal identity, which presupposes uniqueness, leading Parfit to prefer the survival relation over that of personal identity. To summarize, the psychological view is better at dealing with our intuitions when it comes to simple teletransportation cases, but gives poor guidance for personal identity in branching cases, because it allows for multiple realization, which leads to the violation of the premise that personal identity is a one-one relation.

Lastly, we will briefly consider Shapiro's argument against the multiple realizability of the mental, which consequently leads to abandoning

psychological continuity as a criterion of personal identity. Shapiro (2004) argues in detail in favor of the embodiment thesis, or "mind incarnated" thesis, according to which mental states and cognitive processes heavily depend on bodily realizers. This thesis about embodiment is formulated in opposition to standard assumptions about the realizers of cognitive processes and mental states, which traditionally identify neural bodies as an appropriate base of realization. In other words, "mind incarnated" is in direct opposition to "brain theory." But Shapiro adds that the embodiment thesis is also in opposition to the multiple realizability of the mental. Arguments in favor of embodiment are usually based on empirical research, which shows that bodily parts play much more significant roles in many different cognitive processes than previously thought. Shapiro takes into consideration many different examples where processes traditionally thought to be realized in the brain of the subject are heavily dependent on bodily configuration and environmental factors, as well as on relations between the body and the environment. Some of them involve cognitive processes responsible for visual perception (Noë 2005), problem solving (for the famous block copying experiment, see Ballard, D. et al. 1997: 731; Ballard, D. et al. 1995), and linguistic categorization (Lakoff & Johnson 1999). In describing the kind of bodily dependencies needed for auditory acuity, Shapiro writes:

> Generally, larger distances between ears provide greater auditory acuity. But also important is the density of the matter between the ears because sounds of varying frequencies will behave differently when traveling through a given medium. The auditory system incorporates facts about ear distance and head density in its processing, but not in a way that requires their symbolic representation. There is no need to represent the distance between ears because it is the distance itself – not its representation – that creates the opportunity for greater auditory acuity. (Shapiro 2007: 340).

In Shapiro's view it is these kinds of examples that show us the truly incarnated nature of our mind. Concepts of left and right would not make any sense if we were spherical beings; sight wouldn't be possible without the ability to move; and our problem solving abilities would be vastly different without the option of off-loading of the information onto the environment. We agree with Shapiro that cognitive processes and mental states are at least sometimes realized not only in brain matter, but also in other parts of our bodies, as well as in parts of our environment, but we do not agree with Shapiro's view that functionalism and multiple realization are in conflict with the embodied mind thesis. Shapiro reasons that "The claim that minds are multiply realizable suggests that there are no particular physical properties necessary for minds. The claim that minds and bodies are independent, that the properties of the mind can be investigated in isolation

from those of the body, suggests that the mind is like the occupant of a house." (Shapiro 2004: 227) He frames multiple realizability as a separability thesis and argues:

[1] If cognitive processes are multiply realizable then they are separable from their specific realizations and abstractly definable.
[2] Cognitive processes deeply depend on the body in which they are incarnated, and are inseparable from it.
[3] In conclusion, cognitive processes are not multiply realizable. (Milojevic 2013, see also Shapiro 2004, 2007)

In this manner Shapiro is rejecting functionalism as a proper theory of the mental because it cannot accommodate the facts about its embodiment. But Shapiro makes a mistake in claiming that he, like many others, sees "functional, computational, and information-processing approaches to mind as flesh-eating demons" (Clark 2008: 202). Multiple realizability is not supported only by those radical functionalist views that identify mental states with computer programs separable from their physical implementation. This kind of view is also used in the illustration of the modified Brown-Robinson case where mental states are somehow recorded, temporarily stored in a non-biological medium, and then transferred to a different brain, assuming that the bodily differences and brain differences do not matter for the realization of mental states. And for these reasons we believe that such scenarios can be used only as intuition pumps. There are many functionalist positions, especially realizer functionalist positions, which identify mental states and cognitive processes with the realizers of appropriate functional roles. A functional description of mental states does not entail that those states can exist apart from their physical make-up, or at least a certain kind of physical make-up. Chairs are functionally defined, but that does not mean that there could be chairs that are not physically realized. They have to be realized, and they have to be realized in a certain way, namely, so they can afford sitting which will depend on their physical structure. In a similar fashion, functions by which we identify mental states can be realized only by suitable structures, and sometimes these functions will imply a certain kind of physical and chemical structure. Thus, Shapiro's point cannot be applied to all kinds of functionalism.

We believe that multiple realizability of the mental is what gives plausibility and strength to the psychological continuity view. The immediate importance of multiple realizability, and functional determination of mental states, and consequently of psychological continuity, is reflected in the possibility of replacing physical parts in cognitive systems that would not at the same time change their identity. For instance, if a neural implant is used to replace a part of our brain that is malfunctioning, we could include

it in a set of constitutional parts of the self. In more mundane scenarios, psychological continuity secures the identity of the self through bodily changes that happen daily as our cells are being regenerated or replaced, and in a way solves the problem of the Ship of Theseus applied to personal identity. Also, embracing multiple realizability forces us to accept Parfit's conclusion that maintaining personal identity as a one-one relation must reject branching. Although it might seem *ad hoc* to do so, we think that keeping multiple realizability together with a specific view on the realization base is much more important, for the reasons given above, than accounting for a hypothetical branching teletransportation case that can only be resolved by accepting that physical numerical identity is necessary for personal identity. Another important point that we must keep in mind is that, as we saw in the previous sections, we have to be careful about the limits of multiple realization. The brain theory, as a hybrid theory, is both too inclusive and too exclusive. It cannot account for split-brain cases, but allows for brain-transfer cases. Also, if we allow that it is not at all important what realizes mental states, and in doing so admit some kind of separability thesis as described by Shapiro, then we cannot account for the importance of the body in the constitution of the self. This is why we find that a realizer functional ontology of the self, taking into consideration bodily and environmental factors, has the best chance of capturing all that is important for personal identity. The brain theory got the realization base wrong, while computationalism disregarded the importance of the physical properties responsible for the occurrence of mental states. In the sections that follow we will examine the arguments for the embodied and the extended self.

4. Extended and Embodied Cognition

By now it is clear that we are endorsing a kind of functionalist ontology of the self that is careful about the importance of the physical base responsible for the occurrence of higher order functional properties, more precisely, psychological continuity. The hypothesis of Extended cognition is a view motivated by the same kinds of considerations. In its original form, proposed in Clark and Chalmers (1998), mental and cognitive states were described as states that are sometimes externally realized. The thesis is to be distinguished from the content externalism of Putnam and Burge, who differentiated between narrow and wide content, the latter being dependent on the environment. The extended mind and extended cognition theses[6] claim not only that the content of mental states could be widely

---

[6] We talk about two theses instead of one because of different individuation criteria for

realized, but also that mental states narrowly defined as beliefs, desires, etc. could have wide realization. This is why the theses were also dubbed active externalism or vehicle externalism, in contrast to semantic externalism. At the same time, the theses are functionally motivated by the Parity principle[7], and by empirical considerations about the connections between the traditionally recognized cognitive subject and his environment. The functional stance, prominent in the debate about the extension of the mind and cognition, is used to fight the last remnants of Cartesianism. Namely, although many objected to Descartes' postulation of a substance completely distinct from matter, a number of Cartesian assumptions about subjectivity, epistemic attitudes, and the nature of the mind have remained intact. In contemporary philosophy, physicalism became the dominant metaphysical position, but at the same time mind was left to reside enclosed by the boundaries of the skull in the form of a brain, having an immediate and special connection with its own mental contents. Advocating the extended mind thesis means disagreeing with "neural chauvinism," which favors one particular kind of material substance, namely, neural matter. Clark and Chalmers in (1998) offer an argument that should justify the inclusion of non-neural realizers of cognitive processes. The argument can be summarized as follows:

> a) If, as we confront some task, a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognizing it as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process.
> b) There are cases of an external (or partly external) process which functions as a process which, were it done in the head, we would not hesitate to call a cognitive process.
> c) Cognitive processes ain't (all) in the head! (Clark & Chalmers 1998/2008: 222)

The conclusion of the argument, which is at the same time the core claim of the extended cognition hypothesis, follows from the Parity principle, stated by a), and examples provided to support b). The same argument is simultaneously a constitutional argument for the embodiment thesis as a thesis about the bodily realizers of mental and cognitive processes. The

---

mental states and cognitive processes, coming from philosophy of mind and cognitive psychology, respectively.

[7] Some later versions of these theses tended to downplay the importance of functionalism in claiming that sometimes cognitive processes or mental states could, at least partly, be constituted by non-neural matter, as was done by Shapiro (2004, 2007), Menary (2010), and Sutton (2010), and focused on unique realization and the role of transformation and evolution in constitution of cognitive processes. For the purposes of this paper we are going to limit our considerations to functionalist versions of those theses.

argument is, thus, used to turn the evidence about the strong causal dependencies of the brain, body, and environment into an ontological claim about the appropriate realizers based on a functionalist premise. Examples of these dependencies are abundant and can be found in philosophical, psychological, anthropological and neuroscientific literature. (Clark 1997, 2008; Menary 2010; Wilson 2004, Sutton 2010, Damasio 1994, Hutchins 1995, Lakoff and Johnson 1999)

The most quoted example of the extended mind is certainly the case of Otto, stricken with Alzheimer's disease, and his notebook. Although we do not believe that this is the best example that can be found in the literature about extended states, we will describe it, as it will be useful in the remainder of the paper, where we will argue that there are reasons to believe that the constitutive elements of psychological continuity can have wide realization. The case of Otto, his notebook, and Inga is used to offer justification of the premise b) of the Parity argument. It is said that Otto uses his notebook in much the same way as Inga, a healthy cognitive subject, uses her own biological memory. When Otto and Inga are independently told about the particular exhibition happening at the moment in MoMA, they both develop a desire to visit the museum. Alas, Otto cannot rely on his biological memory in remembering where is MoMA located, so he instead consults his notebook. Because the information in Otto's notebook plays the same functional role as Inga's engram in her brain, it is claimed that there are no reasons for not counting the information from the notebook as an instance of a dispositional belief. In addition to the coarse-grained functional roles, such as activation of information after the desire to use it, and subsequent action in accordance with the information, it is said that information must also conform to more fine-grained functional roles later dubbed "glue and trust" conditions in order to prevent overextension. The information has to be readily available, the notebook with the information has to be a constant in Otto's life, Otto has to trust the information in the notebook and automatically endorse it on retrieval, etc.

There are many discussions (Adams & Aizawa 2001, 2008, Rupert 2009) about the appropriate level of functional descriptions of mental states, and whether extended mental states can accommodate them. Because of the scope of this paper we cannot take them into consideration, and we have to limit our goal to focusing on the core insight of the argument. If we take a functionalist stance toward the mind, there are no *a priori* reasons for excluding non-neural matter from the realization base of mental properties. This is not to deny that brain matter plays a crucial part in their realization, but it is a call for broadening this base, especially in those cases where there is a substantial functional gain or restoration of previously impaired

mental and cognitive capabilities enabled by the use of parts of the environment. It is also important to notice that by claiming that "[i]f, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process" (Clark & Chalmers 1998: 8) we are not committed to the claim that there has to be an actual cognitive process which is typically done solely in the head and which functions in a certain way. The claim is restricted only to the point that if it is only the non-neural realization that is keeping us from calling it cognitive, we should refrain from making a negative judgment. This opens the possibility that there may be genuine embodied and extended processes that are contingently never realized solely in the brain. This is the focus of many arguments found in the debate on embodied cognition.

In the literature on embodied cognition we can find a further justification of the claim that the mind should not be taken in isolation from the body and the world. It is often pointed out that much of our cognition that is attributed to the central nervous system as an isolated system is dependent on our bodily shape, its point in space, and our "body schema"[8]. Notably, Gallagher in *How the body shapes the mind* tends to show that our cognitive capabilities are shaped by our bodies in a multitude of ways, primarily by shaping our perception, which underlies our cognitive capacities. In arguing against the reduction of the body to the brain and its representation in somatosensory cortex as a consequence of reducing mental states to brain states (Gallagher 1995), and starting from the insights of some cognitive psychologist such as Neisser (1987) and of the phenomenologists Husserl and Merleau-Ponty, he offers reasons for rethinking the role of the body and bodily experience in explaining cognition. One of the prominent bodily factors that shape our perception and cognition is upright posture, recognized already by Aristotle as an essential part of being human. Upright posture is a trait of our organic body that has a specific structure of the foot, ankle, knee, hip, and vertebral column. This specific structure that enables an upright posture in humans is strongly connected with the state of wakefulness, and the extended range of vision that secures independence, which further shape our perception and enable the development of cognitive abilities. (Gallagher 1995: 147-150) Gallagher, nevertheless, notices that it is not only the shape of the body that constrains perception and action, but also many bodily systems that function

---

[8] "Body schema" is a concept introduced by Head (1920), referring to the nonconsciouss postural model that enables and constrains perception. An important trait of a body schema is that it is constantly updated during body movement.

below the threshold of consciousness, such as metabolism, blood pressure, etc., and that their automatic adjustments affect perception and cognitive performance. Also, by endorsing a Gibsonian (1979) view on perception, Gallagher emphasizes that our bodies structure our perception in yet another way, because our primary way of perceiving is perceiving affordances for the body, and the body schema is responsible for the kind and structure of our interactions with the world. Finally, according to Gallagher the sense of self is also ecologically constituted, and involves a sense of one's own motor possibilities and body posture, as well as the sense of movement and action. Thus, Gallagher shows the importance of the body and our embodiment in explaining cognition as its precondition and limitative factor. A number of authors further show the importance of embodiment in everyday functioning – the role of gesturing in linguistic understanding (McNeill 1992), and problem-solving (Clark 2007), the role of movement and sensory-motor contingencies in visual perception (Noë 2005, Noë & O'Regan 2002), the role of "acting out" in memory (Scott, Harris & Rothe 2001), off-loading into the environment in epistemic actions (Kirsh & Maglio 1994), etc.

While many of these authors explicitly reject functionalism and lean towards new kinds of identity theories, just like Shapiro, we believe that retaining minimal functionalist assumptions is paramount for the reasons we provided earlier. There is no contradiction between embodied and extended cognition and functionalism if we are also careful to specify an appropriate level of functional roles, fine-grained enough to capture appropriate bodily functions, and adopt a kind of realizer functionalism. Multiple realizability is not an enemy of embodiment; it only allows for different types of embodiment. Mental states, cognitive processes, and personal identity can remain the same even if some of their constitutional parts are replaced with those that can play the same functional roles.

So far we have outlined the functional psychological view on personal identity, and how mental states and cognitive processes can be embodied and extended. What we haven't yet established is whether we can talk about extended selves or embodied selves while endorsing both the psychological view and the hypotheses of embodiment and extension of the mental and the cognitive. This will be the topic of our last section.

5. Extended and Embodied Self

Some authors, including Clark and Chalmers, believe that the hypothesis about the vehicle extension of the mental could imply the extension of the self. In (1998) they claim: "Does the extended mind imply an extended self? It seems so. The information in Otto's notebook, for example, is a

central part of his identity as a cognitive agent. What this comes to is that Otto *himself* is best regarded as an extended system, a coupling of biological organism and external resources." (Clark & Chalmers 1998: 18) Others, like Wilson and Lenart (2014), and Lindemann (2009), hold that personal identity does not have to be put in individualistic terms, starting from the psychological continuity view on the self, and that we can also talk about the identity which is collectively constituted. We agree with some of the conclusions these authors make and disagree with others. Arguments for extended mind and cognition can be used to argue in favor of the extended self, and psychological continuity will lead us to consider larger cognitive systems as instantiating personal identity, but limitations have to be imposed on the kind of extended mental states that can lead to the extension of the self, and on the kinds of systems that can be regarded as unique selves.

Namely, if the quoted passage from Clark and Chalmers is taken to entail that any case of extended cognition implies an extended self, then we have to say that the claim is made too hastily. Having one particular extended belief will probably not suffice for the extension of the self. On the other hand, if the same passage is taken to entail only that the case of Otto and his notebook is a case of an extended self, then we would agree with the claims made. But what makes these two interpretations so markedly different? It is the kind of mental processes involved in the extension. Clark and Chalmers say something along these lines when they claim that "[t]he information in Otto's notebook, for example, is a central part of his identity as a cognitive agent," but they do not offer an explanation what makes this information a "central part of his identity." We believe that the hypothesis of extended cognition does not entail the existence of the extended self by itself, but it provides a framework in which we can argue for the extended self if one more condition is met, and that is that the mind is extended in such a way that the basis of psychological continuity is also extended. On the other hand, Wilson and Lenart (2014) correctly emphasize the importance of narrative memory as enabling psychological continuity and the possibility of wide realization of such memories, but by letting the distribution of them onto multiple persons, they violate the assumption that selves are integrated and persist through time.

In order to give a plausible account of the extended self we are going to combine the insights of several authors. Namely, by adopting the psychological account of the self, which is also the focus of Wilson and Lenart, we are going to argue for the extension of the self through possible external realization of relevant memories that are constitutive of psychological continuity. On the other hand, we are going to use Clark and Chalmers'

insights about the nature of constitution of cognitive states, and we are also going to emphasize the role of integration of the processes in the system in considering a cognitive system as a unique entity.

At the very beginning of this paper we said something about the relation between the self and personhood, and we postulated that the self is a unique entity upon which a notion of personhood is based. Having a self is a prerequisite for being a person. In claiming this we are implying that a self is an integrated entity that is determined by criteria for personal identity. Whether a part of the environment can be counted as a part of a self will depend on two conditions: an integration condition and a functional psychological condition. The extended cognition hypothesis gives grounds for arguing that appropriate parts of the environment can fulfill the second condition, as it shows that sometimes parts of the body and the environment should be regarded as constitutive of mental and cognitive states and processes, and Wilson and Lenart (2014) add the important amendment that it is not any extended mental state or cognitive processes that grants the extension of the self but only those that constitute narrative autobiographical memories. The psychological neo-Lockean view on the self does not equate the self with the set of mental states, but with the psychological continuity in a sense that "There is psychological continuity if and only if there are overlapping chains of strong connectedness. X today is one and the same person as Y at some past time if and only if (2) X is psychologically continuous with Y." (Parfit 1984: 207) So the metaphysical basis of the self lies not in the realizers of every mental state, but in the mental and physical capacities that are responsible for psychological continuity. One such capacity is the capacity for narrative autobiographical memory.

Wilson and Lenart focus on narrative memory and empirical research suggesting that this sort of memory gives good guidance in tracking personal identity. They briefly explore "split brain cases," "dissociative identity disorders," and cases of drug induced change of cognitive abilities, as well as cases of voice modulation, and their treatment in the scientific and philosophical literature. All these cases have plausible interpretations if we take narrative memories as a demarcating line of personal identities. Puccetti (1973), Hacking (1995), and Elliot (2003), describe cases of multiple personal identities or selves in one body and cases where there is a "less than one person" in a body, as well as cases of "restoration" of a self, starting from the psychological view on personal identity and the role of personal narratives. Further, Wilson and Lenart argue that this kind of memory is sometimes extended by cognitive offloading, as described in Dennett (1996), Wilson (2004), and Lindemann (2009). The most prominent case of offloading of valuable memories onto the environment that allows for

conducting daily routines is seen in the behavior of elderly people who are tightly connected to their familiar surroundings in their daily functioning. As a second example of extended memories, authors point to the role of collective memories and group dynamics that form narrative memories. Nevertheless, by doing so Wilson and Lenart remove themselves from arguing in favor of extended selves. They explicitly claim that such extensions do not imply the extension of personhood or selves, but still constitute the extension of personal identities. This implies that they will not use the criterion of personal identity for tracking selves, and that personal identity is more broadly construed. We are not going to debate whether such a result points to the fact that psychological continuity is not a good criterion for the identification of the self. Instead we want to show that besides having the relevant function, a part of a self has to be integrated in one unique system. On the other hand, by allowing for distributed mental states, Wilson and Lenart give up on the idea that personal identity can establish personal spatio-temporal boundaries.

Lets go back to Otto and his notebook. Clark and Chalmers claimed that in this case we should consider Otto to be a hybrid system that includes his notebook, which partially constitutes his self. With the addition of the psychological continuity criterion for personal identity, this claim becomes even stronger. Otto's narrative memory and his daily routines heavily depend on the existence of his notebook, because Otto's biological capabilities are deeply affected by his illness and the use of the notebook partially restores them. The notebook is also deeply integrated with Otto's biological body, and this is illustrated in the "glue and trust" conditions the notebook satisfies. The notebook is a constant in Otto's life: he is completely reliant on it, and he uses it with the ease that is characteristic of the use of our biological capacities, meaning that the notebook becomes transparent in its use to Otto. Otto's psychological continuity depends on the usage of the notebook, which functionally satisfies conditions for storing dispositional beliefs or memories. It is also sufficiently integrated into the Otto+notebook cognitive system by further satisfying the trust and glue conditions, thus becoming a true contender for being a part of Otto's self.

While functional psychological conditions are gathered from a scientific discipline, namely, psychology, integration conditions are not so easily spelled out. Conditions of integration may vary according to the kind of the extended processes involved, and different authors put different constraints on the integration. Heersmink (2016), in exploration of the possibility of the extension of moral agency, notices that the integration has to be judged on a number of dimensions. The integration of a non-biological,

non-neural, part of the world into a single cognitive system, one which can instantiate moral agency or a self, will depend on

> The kind and intensity of information flow between agent and scaffold, the accessibility of the scaffold, the durability of the coupling between agent and scaffold, the amount of trust a user puts into the information the scaffold provides, the degree of transparency-in-use, the ease with which the information can be interpreted, amount of personalization, and the amount of cognitive transformation. (Heersmink 2017: 433-4)

This shows that determining whether a part of the environment is sufficiently integrated to be called a part of someone's self is not always an easy task. Nevertheless, there will be indisputable cases of integration; but more importantly, showing the existence of at least one such case gives reasons for arguing that there are extended selves.

6. Conclusions

Starting from the psychological criteria for personal identity, which have fewer ontological commitments than essentialist accounts of personal identity and which can also account for split-brain cases and hypothetical transfer of selves, we have tried to provide reasons for considering selves as embodied and sometimes extended. The main claim of the psychological view was that personal identity is based on psychological continuity, which is secured through "overlapping chains of strong connectedness" (Parfit 1984: 207). One of the main mechanisms for maintaining psychological continuity, and thus personal identity, is narrative autobiographical memory. Taking psychological states and processes to be functionally determined, we differentiated between higher and lower level properties, where higher level functional properties are used for the individuation of these states and processes, whereas lower level properties are responsible for their realization. The functionalism that we endorsed allowed us to talk about multiple realizability, which most importantly made the constitutive base of the self flexible enough to accommodate the persistence of a self even if some constitutive parts of it are replaced and functional isomorphism is maintained.

The novelty of the account we offered consists in broadening the base of realization of the relevant mental and cognitive processes, thus including parts of the body and the environment as constitutive parts of the self. The extended mind thesis offers a strong argument against neural chauvinism, and relies on recent developments in cognitive science showing that the role of the body and the environment are much greater than previously thought in shaping and constituting our cognitive capacities. By rejecting

the claim that it is only brain matter that can be a suitable base of realization of the mental, this thesis dispenses with the last remnants of Cartesianism. Combining the psychological view, functionalism and the extended mind thesis yielded a specific view of the self. The self, or personal identity, on this view, heavily depends on constitutive parts of the self, which are often embodied and can sometimes be extended. Thus, such a view is ontologically more permissive than the hybrid brain theory and animalism by explicitly allowing for non-biological constitutive parts of the self. On the other hand, it puts more constraints on some of the scenarios standardly used for putting our intuitions about personal identity to the test, such are the transfer, transplantation, and teletransportation scenarios. Namely, if it turns out that appropriate functional roles can be satisfied only by physical structures which have relevant functional roles due to their specific topological or physical traits, then it might not be possible to maintain personal identity by transferring only the brain of the relevant subject into a body markedly different from the original one, or to "record" mental states in artificial digital media or the like, only to be transferred into another body. In other words, if the realization basis of relevant functional roles which individuate mental states is wide and includes biological bodies and parts of the environment, then transferring a brain or recorded algorithms implemented in it into a different body would yield a realization of different mental states. In conclusion, the hypothesis of the embodied and extended self is one that makes us rethink the boundaries of the physical realization of our selves. It questions our Cartesian intuitions about the importance of our heads for the persistence of our personal identities. In doing so it combines different intuitions traditionally attributed to the psychological view, but also to animalism. What makes us exist as single thinking, perceiving and acting entities persisting through time might be extended through our bodies and parts of the environment in ways we have not previously imagined.

REFERENCES

Adams, F. & Aizawa, K. (2001). "The bounds of cognition." *Philosophical Psychology* 14 (1): 63-64.

Adams, F. & Aizawa, K. (2008). *The Bounds of Cognition*. Blackwell.

Ballard, D., Hayhoe, M., Pelz, J. (1995). "Memory representations in natural tasks." *Journal of Cognitive Neuroscience* 7: 66–80.

Ballard, D., Hayhoe, M., Pook, P., Rao, R. (1997). "Deictic codes for the embodiment of cognition." *Behavioral and Brain Sciences* 20 (4): 723-742.

Cassam, Q. (2011). "The embodied self." In Shaun Gallagher (ed.) *The Oxford Handbook of the Self*. Oxford University Press: 139-157.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. MIT Press.

Clark, A. (2007). "Curing cognitive hiccups: A defense of the extended mind." *Journal of Philosophy* 104 (4): 163-192.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press.

Clark, A. & Chalmers, D. J. (1998). "The extended mind." *Analysis* 58 (1): 7-19.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. G. P. Putnam's Sons.

Dennett, D. (1996). *Kinds of Minds*. Basic Books.

Elliot, C. (2003). *Better Than Well: American Medicine Meets the American Dream*. New York: Norton.

Gallagher, S. (1995). "Body schema and intentionality." In J. L. Bermúdez, A. J. Marcel & N. M. Eilan (eds.) *The Body and the Self*. MIT Press: 225-244.

Gallagher, S. (2005). *How the Body Shapes the Mind*. Oxford University Press. UK.

Gazzaniga, M. (1995). "Consciousness and the cerebral hemispheres." In *The Cognitive Neurosciences*. MIT Press.

Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Hacking, I. (1995). *Rewriting the Soul: Multiple Personality and the Sciences of Memory*. Princeton University Press.

Head, H. (1920). *Studies in Neurology. Vol 2*. Oxford University Press.

Heersmink, R. (2016). "The Metaphysics of Cognitive Artifacts." *Philosophical Explorations* 19 (1): 78-93.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

Kirsh, D. & Maglio, P. (1994). "On Distinguishing Epistemic from Pragmatic Action." *Cognitive Science* 18 (4): 513-549.

Lakoff, G. & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. Basic Books.

Lindemann, H. (2009). "Holding one another (well, wrongly, clumsily) in a time of dementia." *Metaphilosophy* 40 (4-3): 416-424.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. The University of Chicago Press.

Menary, R. (2010). "The extended mind and cognitive integration." In *The Extended Mind*. MIT Press.

Milojević, M. (2013). "Functionally Extended Cognition." *Prolegomena* 12 (2): 315-336.

Neisser, U. (1987). *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge University Press.

Noë, A. (2005). *Action in Perception*. MIT Press.

Noë, A. & O'Regan, K. J. (2002). "On the brain-basis of visual consciousness: A sensorimotor account." In A. Noë & E. Thompson (eds.) *Vision and Mind: Selected Readings in the Philosophy of Perception*. MIT Press.

Olson, E. (2009). "Self: Personal Identity." In W. Banks (ed.) *Encylcopedia of Consciousness.* Elsevier: 301-302.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Puccetti, R. (1973). "Brain bisection and personal identity." *British Journal for the Philosophy of Science* 24 (April): 339-355.

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

Shapiro, L. (2004). *The Mind Incarnate*. MIT Press.

Shapiro, L. (2007). "The embodied cognition research programme." *Philosophy Compass* 2 (2): 338-346.

Shoemaker, S. (1984). *Personal Identity*. Blackwell.

Snowdon, P. (1995). "Persons, animals and bodies." In Jose Luis Bermúdez, Anthony J. Marcel & Naomi M. Eilan (eds.) *The Body and the Self*. MIT Press.

Scott, C., Harris, R., Rothe, A. (2001). "Embodied Cognition Through Improvisation Improves Memory for a Dramatic Monologue." *Discourse Processes* 31 (3): 293-305.

Sperry, R. (1968). "Hemisphere deconnection and unity in conscious awareness." *American Psychologist* 23: 723-733.

Sutton, J. (2010). "Exograms and Interdisciplinarity: history, the extended mind, and the civilizing process." In Richard Menary (ed.) *The Extended Mind*. MIT Press: 189-225.

Wilson, R. (2004). *Boundaries of the Mind: The Individual in the Fragile Sciences: Cognition*. Cambridge University Press.

Wilson, R. & Lenart, B. (2014). "Extended Mind and Identity." In Jens Clausen & Neil Levy (eds.) *Handbook of Neuroethics*. Springer: 423-439.

3. The Immunological Self

ZDENKA BRZOVIĆ

1. Introduction

The problem of defining the self has traditionally been conceived as a task for philosophers, or to those who are more empirically minded, for psychologists. However, the development of immunology in the second part of the 20th century has led many scientists to conclude that immunology is the science of the self, due to the fact that the role of the immune system has been defined as defending the self from the foreign influence. The discipline grew out of the observation that people who recovered from certain infections were protected against such diseases in the future. One of the main theoretical points of immunology made after the Second World War was the idea that the role of the immune system is to protect the self and attack the nonself, introduced by Burnet (1959) as part of his clonal selection theory.

The problem of defining the biological self belongs to the complex group of issues regarding biological individuality that aim to answer the questions what are biological individuals or agents and what roles they play in various biological processes. Standardly, individual organisms are taken to be exemplary biological individuals. However, this commonsensical view has been questioned, most notably by Richard Dawkins (1976) in *Selfish Gene* where gene is identified as the primary biological individual and a unit of selection. He considered organisms as not stable trough evolutionary time and insignificant in the light of evolutionary theory. In addition, it has been claimed that the concept of organism practically disappeared with the rise of modern synthesis in evolutionary theory and the focus has shifted to the categories of gene and population. (Huneman & Wolfe 2010) Many philosophers of biology have embraced this kind of view and considered organisms as an unclearly defined concept (Wilson 2000) and have argued that the disciplines dealing with organisms such as anatomy and physiology have produced no theories that would inform us how to individuate their objects of interest (Hull 1992).

These kind of considerations led to two different approaches to biological individuality: physiological individuation that is mostly concerned with organisms seen as strongly cohesive and unified metabolic entities, and evolutionary individuation where evolution by natural selection is seen as the best framework for individuating biological entities which are defined primarily as units of selection. (Guay & Pradeu 2015) This paper will be concerned with the physiological approach to biological individuation where organisms are the main category of interest, and more specifically with the question whether immunology can provide us with the criterion of identity for organisms. In the next section I will briefly describe some approaches to organism individuation and point out some of the problems they encounter. Then, I offer an overview of the immunological theories that have been put forward. Special emphasis is put on the self-nonself theory in immunology as the main candidate for providing us with the criterion of an organism's identity. I analyze criticisms directed against this theory and conclude that, it can only function as a very vague metaphor, and not as a real criterion for the identity for organisms in biology.

In continuation, I address some alternative proposals to the self-nonself theory that deny that there is such a thing as an immune self, such as the danger theory that states that the immune system recognizes danger and not selfhood, and the systemic theories of immunity. Finally I will analyze Pradeu's (2012) approach that offers an alternative to the self-nonself theory, a continuity theory and that, in Pradeu's words, can offer a criterion of delineation and individuation of organisms. The aims of the paper are twofold: first is to provide an overview of various immunological theories, their structures, function, and how they approach the question of an organism's identity. Second is to make a conceptual point that all the theories that purport to offer a criterion of identity for organisms fail, because they all presuppose that the identity of an organism is already established and that the job of the immune system is to defend and preserve it.

2. Problems with Defining the Concept of Organism in Biology

One of the main problems regarding organism individuation is whether we can come up with a straightforward criterion of how to delineate individual organisms. In words of Clarke and Okasha (2013) we need a concept of organism that will allow us to count individual organisms, i.e. be able to distinguish organisms from mere parts, and from groups and colonies, and to distinguish reproduction from growth. According to Pradeu, to ask for the identity of some biological entity can imply two different questions:

1) What makes the uniqueness of a living thing, what makes it different from all other living things?
2) What counts as one living being, i.e. what constitutes a discrete, cohesive, clearly delineated unit of the living world? (Pradeu 2012: 2)

Clarke (2010) analyzes 13 different proposed criteria for defining organisms or biological individuals while acknowledging that there are more of them that she failed to include from one reason or other. The main problem seems to consist in specifying how to distinguish organisms from their parts, and from larger groups that consist of larger numbers of organisms. Usually the criteria proposed qualify as being necessary for something to be considered as an organism, but they do not also represent sufficient conditions since many non-organisms can fulfill such criteria as well. (Pepper & Herron 2008) In what follows, I will examine some of the proposed criteria.

Functional integration is a popular criterion for biological organisms that defines organisms as entities that have "evolved to function in a harmonious and coordinated fashion." (Wilson & Sober 1994) The problem with such view is that, unless further specified, it does not offer a clear criterion between organisms and anything else that might be characterized as having a function, being organized so that it produces some function, etc. That is, we can specify traits at different levels (both below and above the organism level) as being functional and functionally integrated. For instance, a cell is functionally integrated if it regulates its internal environment but also some groups of eusocial insects such as bees can count as a functionally integrated individuals. (Pepper & Herron 2008) Thus, unless we have some further criterion specifying what degree of functional integration is required for something to qualify as an organism, we will not be able to use it as a definition of the organism concept.

Autonomy is another popular proposal that also suffers from being vague and underspecified; thus one can try to define organism as being able to sustain itself, being self-sufficient or independent but there is no easy way of measuring autonomy or spelling out what exactly is meant by this. We can identify different levels of autonomy; for instance some authors argue that unicellular constituents of multicellular organisms qualify as autonomous, and that the autonomy at a higher level requires cellular autonomy. (Moreno & Mossio 2015) If one does not posit an additional criterion for establishing a certain degree of autonomy as being crucial for delineating some autonomous structure as an organism, autonomy by itself does not help much in providing a criterion of organism delineation.

The *genetic* criterion delineates organisms by their genotypes either by invoking the uniqueness of each organism's genotype, or at least by invok-

ing the homogeneity of the genotype of the organism's parts (or both). One problem with the genetic criterion is that it leads to many very unintuitive consequences such as delineating a 15 hectares long and 10 000 kg heavy *Armillaria bulbosa* as one fungal organism. Also, identical twins can have identical genotypes but we consider them to be two separate organisms. Another problematic feature with this criterion is that many organisms could not survive without symbiotic organisms (such as for instance gut bacteria in humans), which means that the criterion that invokes genetics can go against a criterion that would specify functional integration as the most important feature.

The *immunological* criterion delineates organisms according to their immune response. According to the self-nonself view in immunology, the organism (or the self) is taken to be everything that is tolerated by the immune system, and the non-self is everything that is attacked by the immune system.

It would be ideal if the proposed criteria could go together so that we can say that we can talk about an organism when all of the enumerated properties coincide. However, this does not happen, and an especially problematic fact is that the genetics criterion does not go together with the criterion of functional integration and autonomy since many organisms cannot survive without (and are integrated with) symbiotic relationships with entities that are genetically distinct from them. Also, there are examples of modular organisms where many physiological individuals are genetically identical but physiologically distinct. (Pepper & Herron 2008) In the next section I examine the self-nonself theory as a proposed criterion of delineating organisms based on the working of the immune system.

3. Self-Nonself Theory

Burnet (1969) characterizes the "immune self" as a lack of response to its own parts (by the organism). His idea was that the basis on which the organism "decides" what to accept and what to reject is genetic. Transplantation is one of the illuminating cases that motivated the self-nonself theory because it appears that with transplantation the organism recognizes its own individuality and rejects everything that is foreign to it, i.e. nonself. "In the first place we have the demonstration that for a tissue to be rejected it must be recognizably different and that the differences involved are genetic in origin." (Burnet 1969: 24)

The main problem is how to characterize the self in a way that can be useful and illuminating as a criterion for an organisms' individuality. If we simply take the self to be that which does not trigger an immune response without further specification we have basically defined self and non-self as

immunogenic and non-immunogenic, and this does not tell us anything further about what is actually going on and why certain immunological reactions occur. From Burnet's original characterization of the self as that which organism's immune systems tolerates (does not attack) it seems that it is already presupposed that there is some further ground for establishing what belongs to the organism and what does not, and then the immune system is somehow able to recognize that. Thus while the theory proclaims to define or delineate the self, it appears that something else must be the basis for determining the identity of the organism, and the immune system is just the capacity to recognize that entity.

As stated, Burnet defines the organism's identity genetically, but the problem with this kind of approach is that it does not account for the fact that a mother's organism tolerates the fetus (does not reject it) even though half of the fetus' genetic material is foreign. Also, it does not account for the commensal bacteria living inside our bodies tolerated by our immune systems. It has been established that genetically foreign matter introduced to an organism during fetal development is tolerated by the organism throughout its lifetime. Thus, there are well established cases of organisms tolerating genetically foreign biological antigens.

There have also been attempts to define the self phenotypically, but this approach cannot explain why the organism does not reject transplants from genetically identical individuals. A more specific immunological approach was to identify the self with the organism's tissues - all markers of the major histocompatibility complex (MHC), or HLA in humans as organisms' molecular "identity card." The MHC consists of cell surface proteins which bind to pathogens' peptide fragments and present them on the surface of the cells so that the immune system is able to recognize them. MHC is important because it allows the immune system to recognize and not react against itself. Also, it determines whether a donor of organ transplant is compatible with the host. A very high degree of diversity of molecules of the histocompatibility system allows these molecules to be treated as manifestations of organism's uniqueness and one of the best ways for delineating biological individuals. (Pradeu 2012) This way of establishing an organism's identity explains the cases of transplant reception and acceptance. However, it does not explain already mentioned cases of immune tolerance such as the fact that the fetuses' tissues are tolerated by the mother's immune system.

Regardless of whether we can specify one specific delineation criterion for the self that the immune system is supposed to defend, there seem to be problems in the self-nonself theory that put in doubt the whole idea that the immune system protects the self from the foreign. Thus, even if we

accept that the self in question is a vague and underspecified concept, the theory has further problems that bring into the question the discourse of the self versus nonself.

A general problem with the self-nonself theory already noticed by Burnet is how to account for the phenomena of autoreactivity, the fact that the immune system checks the organism for some abnormal endogenous modifications (mutations due to age, cancer, etc.) and reacts against such "self" components. Thus, immune cells can in certain circumstances react to antigens created by the genetic self. I should add that various forms of autoreactivity that are considered pathological such as autoimmune diseases are not considered as problematic since the self is defined by the immune system of an organism that is functioning properly.

Normal autoimmunity is what allows an organism to maintain homeostasis. Pradeu (2012) criticizes the self-nonself theory for its inability to distinguish between autoreactivity, autoimmunity and autoimmune disease. Autoreactivity occurs in all cases where an immune cell's receptors interact with an endogenous antigen, and autoimmunity refers to situations where immune activation, i.e. either the destruction of the target or inhibition of distraction against an endogenous antigen is triggered. Autoimmune disease, on the other hand, is when an organism's immune system triggers a destructive response against its own tissues. Thus, normal autoreactivity comes down to the fact that the immune system constantly surveilles all the organism's components, and if the immune cells are activated, destruction ensues. According to Pradeu (2012) some form of such immune surveillance exists in all plant and animal species and it exists in both central and periphery organs. Thus, not only does the immune system react to "self" elements, but such reactions are necessary for maintaining it in a healthy state. Normal autoimmunity also occurs in every organism, when endogenous antigens trigger immune reactions. For instance, phagocytic cells play the role of "garbage men" and get rid of organisms' waste, such as dead cells. However, one must be careful to distinguish normal autoimmunity from autoimmune disease. While autoreactivity and autoimmunity both precede autoimmune disease, the autoimmune disease constitutes a dysfunction in autoreactivity and autoimmunity.

Another large problem for the self-nonself theory is the phenomenon of immune tolerance, the absence of immune response against some foreign or nonself entities. According to Burnet (1969) and his followers this was seen as an exception to the rule of self-nonself discrimination and it referred primarily to the period of immune immaturity found in many animals in which the presence of foreign components is tolerated because the organism is still learning the ability to recognize its own from foreign

components. However, the phenomenon of immune tolerance is not limited to the period of immaturity and it occurs quite often. It can refer to two things: immunosuppression - a non-rejection of a genetically foreign entity due to the absence of appropriate immune cells, or immunoregulation – the inhibition of destructive immune response due to activation of certain immune cells. The tolerance that is of interest in this paper is immunoregulation. Pradeu (2012) analyzes two important phenomena where it is obvious that the organism does not reject any foreign entity: fetomaternal tolerance and the tolerance to macro and microorganisms.

When we talk about graft tolerance, it is usually taken that the organism tolerates grafts taken from their own tissue, or those from genetically identical twins. However, there is the case of immunoprivileged organs that, when transplanted from one organism to another, do not trigger immune response, or trigger a very weak one. Such properties are found in cornea, brain (and perhaps the entire nervous system), testicles, and the fetus that does not trigger an immune reaction from the mother's organism. When such phenomena were discovered, it was hypothesized that those sites are somehow isolated from the immune systems' action, but more recently it was discovered that mechanisms that allow immune privilege are very similar to the ones that provide tolerance to the commensal microorganisms (Mellor & Munn 2008). In the case of pregnancy, as already mentioned, the self-nonself theory is also facing problems because it cannot explain why the mother is not rejecting the fetus. An earlier explanation for this was that placenta acts as an impenetrable barrier to immune cells, but it was shown that immune actors are present in the placenta but do not trigger a reaction. Fetomaternal tolerance is an active phenomenon that involves immune components that play roles in other tolerance processes. Also, after giving birth the mother conserves cells from the infant she carried for a very long time, even for the entire lifetime. This phenomenon has been called fetomaternal chimerism.

The last very important finding that goes against the self-nonself theory is the tolerance of commensal and symbiotic microorganisms. Symbiosis is a lasting relationship between two organisms belonging to different species that is beneficial for at least one of them. Humans are in symbiosis with numerous bacteria: on the skin, in the intestine, lungs, vagina, etc. Also, most multicellular organisms contain a large number of symbiotic bacteria. For instance, symbiotic intestinal bacteria are unique to each organism, and can constitute one of the best ways to individualize the organism. It was previously considered that that the immune system had no access to intestine bacteria, but this has proven to be untrue. Thus, all the evidence points to the fact that immune tolerance is not about exceptions, but a frequent

and regular part of the immune system's activities. According to Pradeu (2012) the immune system evolved in the dual direction; a capacity to destroy target elements, and a capacity to regulate this destructive response to avoid excessive damage.

All the phenomena examined demonstrated that it is not the case that the organism tolerates the self and rejects the nonself. Rather, each organism is very heterogeneous and it is populated by a vast number of foreign entities like bacteria, viruses and parasites. Thus, according to Pradeu (2012) it is wrong to conceive of an organism as perfectly homogenous and endogenously constructed entity.

To recapitulate, the self-nonself theory either suffers from the problem that the self and nonself are too vaguely defined so that the theory does not explain much, or when it is specified by referring to organisms' genome or phenotypic properties the rule of rejection of nonself and tolerance of self does not hold. Because of such reasons, Ana Marie Moulin (1990) and Alfred Tauber (1994) both talk about the self as a metaphor. The self-nonself theory does not, in the present form, establish a criterion for organisms' identity. In the next section I will briefly examine the systemic theories of immunity that arose as criticism of the self-nonself theory.

4. Systemic Theories of Immunity

Systemic theories begin with Jerne (1974) who sees the immune system as self-centered and autoreactive because it mostly or only deals with the self. This is due to the fact that an organism's antibodies are initially produced in the absence of nonself and they continue to interact with the organism's components, i.e. they do not react to environmental antigens but rather to how the immune system's antibodies express certain antigens. The basic idea is that any immune reaction comes down to a kind of autoreactivity, and the foreign entities provoke an immune response only if they bring about a disturbance of the immune system. Jerne tries to draw parallels between the immune system and the nervous system, and present the immune system as a kind of cognitive agent.

One of the ideas connected with Jerne's view is the theory of autopoesis introduced by Maturana and Varela (1980) that claims that the organism is self-constructed and autonomous. The key claim is that the organism cannot be influenced from the outside because it has to interpret any information or entity before it interacts with it. Thus, we cannot talk about the self and nonself in immunity because there can only be the self, all the interactions are directed inwards: "(…) the organism perceives the penetration of foreign materials not by recognizing them as foreign, but rather because the foreign materials interfere with ongoing reactions which exist

as links in a complex network of interactions. The organism responds to an 'internal image' of the foreign molecule, to its meaning translated in terms of the language previously utilized by the network." (Vaz & Varela 1978: 251, 252)

The main problem with this type of views is that they are vague so that it is not entirely clear what the main contribution consists in. If what is meant by the autopoesis is that exogenous antigens are processed by the organism before encountering the immune system, then the claim is unproblematic, but is does not bring anything new to the table. On the other hand, if the claim is really that the immune system deals only with its own components then it seems that this is not true. Thus it appears that the theory offers very few testable hypotheses that could be experimentally examined. In what follows I will examine the danger theory of immunity that states that the immune system reacts against the danger, and not non-self.

5. The Danger Theory

Polly Matzinger (1994) proposed the "danger theory" according to which the immune response is initiated by the fact that the immune system recognizes the substance as dangerous. Instead of discriminating between the self-components and foreign components, it reacts to inflammatory signals from the damaged cells or the ones that are dying abnormally. Thus, the self constituents that are recognized as dangerous can trigger an immune response, and non-self constituents such as commensal bacteria can be tolerated because they are not recognized as dangerous. For example, the fact that the mother's immune system does not reject the fetus is explained by the fact that the fetus does not represent danger to the mother, or commensal bacteria is tolerated because it does not threaten the organism containing it.

The main issue when examining the danger theory is to explore whether it offers a more testable criterion of what counts as danger at the immunological level and how the immune system recognizes that something is dangerous. The idea is that the damaged or abnormally dying cells produce some sort of a signal that the immune system is able to differentiate from the signals coming from the normal cells. But why are transplanted organs rejected by the immune system if they are not dangerous (rather they are beneficial) for the organism?

Matzinger's response is that the danger signal is produced by the injury that is a consequence of the surgery. She thinks that the danger model can explain why some transplanted organisms are rejected while the others are tolerated even though both constitute the non-self. For example liver transplants are tolerated much more often than skin or heart transplants

because livers can regenerate new cells after they are damaged. The tissue damage that resulted from the surgery will induce the immune response at the beginning but the regenerating cells that will be produced afterwards will be tolerated. With time, the damaged cells that provoked the immune reaction will die out and no longer produce distress signals, and the liver can survive the attempts of rejection and develop tolerance through new regenerated cells that do not produce distress signals.

However, the definition of danger still remains unclear, what does it mean that the cell dies abnormally? For example, Pradeu states that a cell can die from necrosis without causing damage to the organism's tissues. (Pradeu 2012) Thus, unless the theory is able to specify exactly how these danger signals are realized at the molecular level, it remains too vague. If it is specified that the danger actually refers to the damaged tissues then again it ought to be specified how exactly a damage signal is recognized at the molecular level. For example, one can argue that an organism's commensal bacteria are tolerated because they are not dangerous, but most of the pathogenic bacteria rejected by the immune system get rejected prior to causing any damage, so it is not clear what is it that the immune system recognizes as dangerous. (Pradeu 2012) Also, there can be immune response without tissue damage, for example grafts achieved without any damage or inflammation have been noticed to trigger a strong immune response. (Bingaman et al. 2000) Thus, it appears that the danger theory is not able to offer a clear testable criterion for what the immune system rejects. Pradeu (2012) offers an alternative to the self-nonself theory that will provide a criterion for organisms' identity. I will examine his theory in turn.

6. Pradeu's Continuity Theory

This theory states that an immune response is triggered when there is a strong modification of the antigenic patterns with which the organism's immune receptors interact, i.e. a sudden appearance of very different patterns form the ones that the organism is usually interacting with. Those patterns can both be endogenous and exogenous, and the difference is between antigens that are present long-term and those that appear suddenly. Only a strong antigenic discontinuity is immunogenic on this proposal, since not all the expressions of unusual patterns trigger an immune rejection. The main point of the theory is to offer a definition of strong discontinuity that will be specific enough to allow us to draw some conclusions about when an immune reaction will or will not occur. This is done in molecular terms; molecular patterns that are constantly present in the organism and in the interaction with immune receptors do not activate im-

mune response, while those with an unusual molecular pattern lead to an immune response. Thus, antigens that differ substantially from those with which the organism normally interacts are the ones that cause the reaction. Foreign patterns do very often trigger a response, but not because they are foreign but because they are strongly different from the ones the immune receptors usually interact with.

The main problem is to determine at what point in time the constant and repeated activity of antigens establish their continuity. The immune system develops in the early embryonic period but it still continues to develop even after birth. Thus, the continuity that the continuity theory refers to begins with the maturation of the immune system. In what follows I will state some of the more specific factors of continuity and discontinuity as stated by Pradeu (2012):

> 1. Antigen quantities – if the quantities of antigen are small they will most likely not produce an immune response. On the other hand, very large quantities of antigens might paralyze the immune system. Small quantities of antigen presented repeatedly to immune cells can bring to a state of immune tolerance.
> 2. The speed of antigen appearance – antigens that appear progressively do not provoke immune response while those that appear suddenly will provoke it.
> 3. The degree of molecular difference – between antigens that are constantly presented to immune cells and those that appear at a given moment.
> 4. The regularity of antigen presentation – if the antigen is regularly present it can lead to immune tolerance.
> 5. The site of immune reaction – the location of reactions between immune receptors and ligands is relevant for determining whether a reaction will occur. Thus, an antigen that is tolerated at one location (for example in the intestine) can trigger a reaction if introduced to some other location in the organism.

The points 2, 3 and 4 account for the fact that the organism has normal endogenous reactions – throughout its lifetime the organism changes and its tissues are modified, it undergoes mutations that can even have phenotypic consequences. Due to the aforementioned factors, the changes in question get incorporated into the organism's immune systems because the changes in antigens are most often very similar to the organism's usual antigens, they appear slowly or are repeatedly presented to the organism's immune receptors. In cases where these conditions are not fulfilled an immune reaction against the organism's own components is triggered. This is exactly what the theory states – that strong antigenic discontinuities, whether

exogenous or endogenous trigger an immune response. For example the growth of tumors is one of the cases where such a situation occurs.

Pradeu states that the continuity that his theory invokes is spatiotemporal because the place of the immune reaction is relevant, and time is important due to the fact that the speed of antigen's appearance and the regularity of interactions is relevant for the occurrence of immune reaction. It is also important to note that the theory does not state that once the period of mature immunity is reached, everything that is alike to the existing state will be conserved. In contrast to the self-nonself theory that takes the organism as closed to the environment with some exceptions to this for special cases, the continuity theory takes the organism to be primarily open to the environment, but it can reject entities if they are harmful to it. Pradeu thinks that his theory can offer a criterion for the organism's identity because, unlike the criterion of functional integration which is vague because there can be different degrees and levels of functional integration, immunity is organismic, i.e. they concern the whole organism. (Pradeu 2010) Also, immune reactions can provide us with a criterion for distinguishing what constitutes an organism; so all the continuous molecular components that are not rejected by the immune system can be said to belong to the organism. This is his proposed definition of an organism:

> An organism is a functionally integrated whole, made up of heterogeneous constituents that are locally interconnected by strong biochemical interactions and controlled by systemic immune interactions that repeat constantly at the same medium intensity. (Pradeu 2010: 258)

This means that an organism is composed of heterogeneous parts that could have come from the "outside" (parts that have not originated in the organism) such as various bacteria that play physiological roles in the organism, and that can have different genetic makeup. The main difference between this view and the self theory is the fact that the self theory identifies the self with endogenous components, while Pradeu dispenses with the notion of the self, and considers the organism as a whole constituted of heterogeneous parts that are kept together through functional integrity and the action of the immune system. However, he states that his definition does not imply that everything that does not trigger an organism's immune response actually belongs to this organism because we can have the case of identical twins where a transplanted organism from one twin is accepted by the other's immune system, but it does not follow that they are one and the same organism. In order for a part to belong to an organism it must be functionally integrated inside that organism and not rejected by the immune system.

It appears that Pradeu's theory manages to avoid some of the problems facing the self-nonself theory and the theories that followed it. However, the question remains whether it can solve the main problem facing the self-nonself theory if it wants to offer a criterion of an organism's identity and not just a testable criterion of how the immune system works and what it reacts to. All the theories examined, while differing in their success of accounting for specific immune phenomena, seem to be unable to provide a criterion of what makes an organism's identity. Rather, they seem to presuppose it, and the immune system plays a role of preserving that identity. Continuity theory is not different in this sense, it states that what gets accepted are the molecules and tissues that are continuous with the ones already belonging to the organism, but this does not provide the criterion for the organism's identity, it merely states that the immune system preserves (and perhaps helps us delineate) the existing identity of the organism. This is clear from the claim that in order for a part to belong to an organism it must be functionally integrated inside that organism. Thus, it would appear that it is functional integration, and not the immunity that provides us with the criterion for an organism's identity. One could say that the working of the immune system can serve as a useful tool of establishing the boundaries of the organisms whose identity is already presupposed.

However, this objection is not specific to Pradeu's theory, all the immunological theories, even the self-theory presuppose the identity of the organism that the immune system is protecting. Even the views that are exploring the first occurrence and evolution of the immune system presuppose that the proto forms of immune system were evolved in order to protect an already formed entity (even if very unstable one). Thus, one might conclude that the whole project is phrased too ambitiously in the first place. But theories such as Pradeu's can still be of help in delineating organisms. We saw in the first part of the paper that the criteria for an organism's identity such as the functional integration that Pradeu invokes are not offering a clear criterion as to where to draw the line between different degrees and levels of integration, and the continuity theory gives us a testable criterion, an organism is a functionally integrated whole that is joined together by the protection and surveillance of the same immune system. The empirical claims that the theory makes remain to be tested, but its main advantage when compared with other theories of immunity is that it offers a testable criterion for the delineation of organisms.

7. Conclusion

I addressed the various theories in immunology and examined if they are able to provide us with a possible solution to the issue of an organism's identity. A special focus was put on the self-nonself theory, and the criticisms made against it. Furthermore, I examined some alternative proposals to the self-nonself theory that deny that there is such a thing as an immune self, such as the danger theory that states that the immune system recognizes danger and not selfhood, the systemic theories of immunity and Pradeu's continuity theory. Continuity theory is interesting because it explicitly claims to offer a criterion of an organism's identity, but I argued that at best it can be taken as a useful tool for establishing the organism's boundaries, while something else (in the case of the continuity theory this is functional integration) establishes the organism's identity. However, this is a problem for all the proposed theories of organism individuation that rely on immunological criteria – they already presuppose the existence of the organism that the immune system then defends. The main upshot of the continuity theory is that it offers a testable tool for establishing what level of functional integration counts as an organism, namely parts that are functionally integrated as defended by a common immune system.

REFERENCES

Bingaman, A. W. et al. (2000). "Vigorous allograft rejection in the absence of danger." *Journal of Immunology* 164: 3065–3071.

Burnet, F. (1969). *Self and Not-Self: Cellular Immunology.* Cambridge University Press.

Clarke, E. (2010). "The problem of biological individuality." *Biological Theory* 5 (4): 312-325.

Clarke, E. & Okasha, S. (2013). "Species and Organisms: What are the Problems?" In Humenan, P. & Bouchard, F., *From Groups to Individuals. Evolution and Emerging Individuality.* MIT Press: 55-77.

Dawkins, R. (1976). *The selfish gene.* Oxford University Press.

Guay, A. & Pradeu, T. (2015). "Progressive steps towards a unified conception of individuality." In Guay, A, & Pradeu, T. *Individuals Across the Sciences.* Oxford University Press: 1-25.

Hull, D. (1992). "Individual." In E. Fox Keller, *Keywords in Evolutionary Biology.* Harvard University Press.

Huneman, P. & Wolfe, C. T. (2010). "The Concept of Organism: Historical, Philosophical, Scientific Perspectives." *History and Philosophy of Life Sciences* 32: 147-154.

Jerne, N. K. (1974). "Towards a network theory of the immune system." *Annales d'immunologie* 125 C: 373–89.

Maturana, H. R. & Varela, F. J. (1980). *Autopoesis and cognition.* Dordrecht: D. Reidel Publishing Company.

Mellor, A. & Munn, D. (2008). "Creating immune privilege: active local suppression that benefits friends, but protects foes." *Nature Reviews Immunology* 8 (1): 74-80.

Metzinger, P. (1994). "Tolerance, danger, and the extended family." *Annual Review of Immunology* 12: 991–1045.

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

Moulin, A. M. (1990). "La Métaphore du soi et le tabou de l'auto-immunité." In J. Bernard, M. Bessis & C. Debru *Soi et non-soi.* Paris: Seuil: 55–68.

Pepper, J. & Herron, M. (2008). "Does Biology Need an Organism Concept." *Biological Reviews*: 621-627.

Pradeu, T. (2012). *The Limits of the Self.* Oxford University Press.

Pradeu, T. (2010). "What Is An Organism? An Immunological Answer." *History and philosophy of the life sciences*: 247-267.

Tauber, A. (1994). *The Immune Self: Theory or Metaphor?* New York: Cambridge University Press.

Vaz, N. M. & Varela, F. J. (1978). "Self and non-sense: an organism-centered approach to immunology." *Medical Hypotheses* 4 (3): 231-67.

Wilson, D. & Sober, E. (1994). "Reintroducing group selection to the human behavioral-sciences." *Behavioral and Brain Sciences* 17: 606.

Wilson, J. (2000). "Ontological Butchery: Organism Concepts and Biological Generalizations." *Philosophy of Science* 67: 301-313.

Part II

SELF–KNOWLEDGE

4. The Value Of Self-Knowledge

NENAD MIŠČEVIĆ

1. Introduction

The topic of self-knowledge, and self-examination geared to acquiring it, including the issue of the value of the two, is one of the oldest in philosophy; on the other hand, it has experienced a blossoming within analytic tradition in the last half century or so. Our Rijeka conference bears witness to it.[1] The present-day discussion in fact combines several areas: on the side of epistemology, we have the general issue of the value of knowledge and of its origin (assuming knowledge has a value). On the intersection of epistemology and philosophy of mind we have the topic of self-knowledge, its character and origin. Finally, on the side of traditional philosophical interest in good and meaningful life, reaching into the area of ethics, we have the topic of wisdom and the examined life.[2]

We shall assume, in agreement with the vast majority of philosophers who have written on the topic, that self-knowledge has some value. Our central question will concern its character. Is it intrinsic or instrumental or both? The alternative has been, I think, implicitly present since the very beginning of philosophical investigation into the topic. Take Socrates and his famous claim that "the unexamined life is not worth living." (*Apology* 38a) The examined life is a life of knowing oneself, recommended by the oracle, who seems to recommend to us a life of permanent self-inquiry and an attitude of permanent self-inquisitiveness. On the most popular reading, the value of examined life and of knowing oneself has a high instrumental nature: this knowledge will make one morally more capable, prudentially more successful, and the like. But in another famous place, the *Apology* offers a slightly wider picture; here Socrates is asking the Athenians whether they are not ashamed for caring only for "having as much

---

[2] See (Miščević 2012a) and (2012b).

money as possible, and reputation, and honor," but not caring or giving thought to "prudence[3] and truth, and how your soul will be the best possible." (*Apology* 29 d8-e3)

A few lines before (at 29b), he suggests that "the most reprehensible form of ignorance is that of thinking one knows what one does not know." One might read these passages as suggesting that knowledge and truth do have some value in themselves; in the given context that would imply that knowing the truth about oneself is clearly intrinsically valuable.[4]

One speaks of self-knowledge at different levels. Let me just remind you of some we will not address here. First, the self-location level, at which we locate ourselves in the world. Where am I? What date and time is now? as the suddenly awakened Sleeping Beauty might ask. Or, more ambitiously, which of the two gods am I? as Castor wonders in David Lewis's thought experiment (1979). Second, the level of knowledge of one's internal states. First, immediate experience. Next comes semantic self-knowledge, and its problematic implications for externalist views: if water-thought refers to H_2O in virtue of its causal links to some actual H_2O in the surroundings, and if I can know from the armchair that I have a water-thought, than I can know from the armchair that there is H_2O around, which seems strange. And the list of levels and kinds goes on.

In this paper we will not talk about all these kinds of self-knowledge; rather, we will discuss two typical and very distinct kinds. On the one hand, we will briefly discuss the knowledge of inner phenomenal states, such as my knowledge that I feel pain in my back, and on the other hand knowledge of one's causal and dispositional properties (CD-properties, for short), such as my knowledge that I am a gourmet, or that I am prone to jealousy. The choice is easy to justify: the first kind is probably the one most discussed in the mainstream analytic literature on self-knowledge, whereas the second has been the focus of traditional and tradition-inspired reflection on self-knowledge, from the ancient Greeks to Foucault, and is again prominent in the work of authors such as Qassim Cassam. In this sense the two kinds are paradigmatic articles of epistemic evaluation.

---

[3] "Phronesis" Grube translates as "wisdom."

[4] On the point of caring about truth (of one's beliefs), see the summary of proposed readings in Christopher Rowe's (2011) chapter on Self-Examination in *The Cambridge Companion to Socrates*. On the general topic of self-examination see also Kraut, Richard (2006), "The Examined Life" in Sara Ahbel-Rappe and Rachana Kamtekar (eds.), *Blackwell Companion to Socrates*, Blackwell 2000.

A much more radical, but in some respects limited line in favor of the intrinsic value of knowing oneself is present in the Aristotelian tradition, with the ideal of the Intellect thinking (about) oneself; it is not a kind of self-knowledge we would think of today, but it is worth mentioning.

Here is the plan. In the next section we first present the two kinds of self-knowledge in some detail, and then pass to our central question of value, distinguishing four cases, resulting from the combination of kinds of knowledge and the two kinds of epistemic value, extrinsic-instrumental and intrinsic. Section Three is dedicated to polemics about the higher and/or intrinsic value of the knowledge of one's long-term dispositions and causal powers, with authors like Simon D. Feldman, Allan Hazlett and, above all Cassam, who deny such value to it (see References). We shall connect the intrinsic value of higher kinds of self-knowledge to the epistemic virtue of self-inquisitiveness, thus briefly indicating the place of our discussion within the framework of virtue epistemology.

2. The Big Picture

2.1. Varieties of Self-Knowledge

We have already mentioned two articles to be evaluated, the two kinds of self-knowledge that will interest us here: immediate phenomenal knowledge of occurrent episodes and knowledge of one's causal and dispositional, (CD) properties. For knowledge of phenomenal occurrent episodes, take the following as an example: when Mary the neuroscientist sees what she "never had seen," she famously learns what it is like to have the experience of red. (The self comes in indirectly, with the question of how internal the matters are, and with the radical externalist denial that there is something fundamentally internal to them.) The second kind has to do with one's causal powers, active and passive. One causes things, acting in the world, and also, acting on oneself. Thus, causal-dispositional-level concerns, causally oriented active and passive dispositions of one's self, and thereby the ways of being (possibly) affected by various courses of things, and of reacting to them: *reliably true factual beliefs about causal structure in human matters*. Such knowledge of a causal-dispositional sort can be obtained from a variety of sources: experience, introspection, simulation (including thought experiments), psychology, psychoanalysis and so on. John Campbell claims that self-consciousness

> …involves grasping one's own causal structure. There are two dimensions to this grasp of causal structure. There is grasp of the idea that one s later states causally depend on one's earlier states … The other dimension in grasp of one's own causal structure is the idea that one can function as a common cause of various correlated events around one. (Campbell 1994: 2)

Of course, this sort of "self-consciousness" includes self-knowledge of one's immediate states as its component (I don't just feel pain, but I am aware that it comes from a particular source), and meshes well with self-knowl-

edge of causal causal-dispositional sort, which just is "grasp of one's own causal structure" as Campbell would put it.

Phenomenal knowledge certainly is paradigmatic for one genus of self-knowledge. But why did I chose CD-knowledge as a counterpart on the opposite, more objective side? Because it has a long and respectable history that makes it paradigmatic. In a perceptive comment, Anthony Hatzimoysis has observed that "for the ancients self-knowledge is primarily a good to be achieved, whereas for the moderns it is mainly a puzzle to be resolved." (in his Introduction to *Self-knowledge*, 2011: 1) I surmise that the selves to be known are not at the same level: the targeted level for the ancients is the causal-dispositional structure of one's self, whereas for the moderns the target is the more immediate kinds. In order to be(come) wise I have to know my motives and my habits, my ways of reacting to external events, opportunities and pressures, and about the methods that could change these ways. "Knowing what we are, we shall know how to take care of ourselves, and if we are ignorant we shall not know," writes Plato. (authentic or not, in *Alcibiades* 129a, Jowett translation)[5] So much about first-level knowledge in general.

So, we have two kinds of self-knowledge to discuss. Each belongs to a separate, wider genus: the first is a kind of direct knowledge of inner states, and the second a kind of objective, mostly inferential knowledge. The contrast between the two genera (plus or minus some small differences) is often noticed in the literature.

Cassam has value-laden terms for the contrast, and speaks of trivial vs. substantial self-knowledge. (Cassam 2014: Ch.3) His examples of the later include knowledge of one's own character, values, abilities, aptitudes, and emotions, plus knowledge of what makes one happy and why one's attitudes are as they are.[6] My contrast is less evaluative: I would count all knowledge

---

[5] This is why the proverbial wisdom is so often couched in explicit or implicit conditionals, prefaced or accompanied by the point of the conditional, like e.g. "You should not vouch for someone: that man will have a hold on you" *The Instructions of Shuruppag*, and hundreds of others. Katharine J. Dell writes: "The thought-world of *Proverbs* is that of "the act–consequence relationship"(…), that is, the principle that good and bad deeds have consequence that can be known through the study of patterns of human behavior. This principle and various other insights into human characteristics are summed in pithy proverbial sayings, the fruit of the experiences of many generation." (2001: 418) For an overview, see Dell 2011.

[6] See the list of conditions that make such knowledge substantial on pp. 31 ff. of his (2014) book. Let me just mention a few; the names speak for themselves:
(i) The Fallibility Condition, (iv) The Challenge Condition, (v) The Corrigibility Condition: (vi) The Non-Transparency Condition, (viii) The Cognitive Effort Condition and (x) The Value Condition.

of my dispositions as cases of CD-knowledge (e.g. of one's preference for a particular kind of ice-cream, which he would not count is as substantial)

Annalisa Coliva in her *Varieties of Self-Knowledge* (2016) comes closer to the present contrast of the two genera. Hers is a first-person vs. third person contrast, with no preferences and evaluations, at least officially. In Chapter two of her book, she talks on the one hand about "truly first-personal self-knowledge" and on the other about "third-personal self-knowledge." The first kind is characterized by groundlessness, transparency and authority (§1), properties that are "not contingent but necessary and a priori aspects of what goes by the name of 'first-personal self-knowledge'." (2016: 58) However, it is important to note that CD-self-knowledge, though lacking groundlessness, transparency, and authority and thus not counting as *truly* first personal, still involves a self-conscious reference to the thinker herself: "I am prone to jealousy" is neither groundless, transparent nor specially authoritative, but still involves the fact that it is a thought attached to the thinker in a special way (see the remark on Perry's idea of self-attached knowledge below).

So much for the two recent and congenial classifications. I shall not talk much about the first kind, inner-state phenomenal knowledge; it has been endlessly described and discussed in the literature. Since I am not interested here in its metaphysical status, which is the biggest issue in the literature, but only in its epistemic value, I can accept the agreed phenomenology and then pass directly to the issue of the value.

Let me say a bit more about my second article for evaluation, CD self-knowledge. We start with a simple and realistic example. Sitting at the computer, I feel pain in my lower back and I change my position. First, the feeling of pain is brought into connection with my posture. Second, I know (at a very elementary level) that the pain will stop (or get less intense) if I change my posture. Pain-posture-changes are part of a causal structure implicitly known by the agent, and this implicit causal knowledge is a pre-requisite for action (as has been stressed by Perry, Campbell, and Damasio).

Suppose that again I feel pain. But I feel it as-coming-from-an-object, for instance a dog's teeth; I was playing with my daughter's dog Kira, and I overdid it. The object, Kira, is affecting me. I withdraw my hand. Again, I am reacting on the basis of expectations of causally organized course(s) of events, this time starting from me. Such simple causal connections to myself are there, presented in most elementary self-knowledge of the CD variety. Its core is the grasp of one's own causal structure, as J. Campbell would put it. For him, self-consciousness:

> …involves grasping one's own causal structure. There are two dimensions to this grasp of causal structure. There is grasp of the idea that one's later states causally depend on one's earlier states … The other dimension in grasp of one's own causal structure is the idea that one can function as a common cause of various correlated events around one. (Campbell 1994: 2)

I shall not call it self-consciousness here, but rather self-knowledge. Non-human animals might enjoy it in some primitive form. Jose Bermúdez, for example, argues that "bodily awareness is a basic form of self-consciousness, through which perceiving agents are directly conscious of the bodily self." (2012: 157), and notes its "immediate implications for action" (2012: 168). In his "Summary of *The Paradox of Self-Consciousness*" he gives a principled formulation:

> The nonconceptual first person contents implicated in somatic proprioception and the pick-up of self-specifying information in exteroceptive perception provide very primitive forms of nonconceptual self-consciousness, albeit ones that can plausibly be viewed as in place from birth or shortly afterwards. (no pagination, available at http://host.uniroma3.it/progetti /kant/ field/ Bermúdezsum.htm)

So much for the primitive level. The deeper the agent goes in self-locating and self-concerning thought, the richer and more interesting the causal structure gets. I can start from knowing about the position of my hand, and about my occurent state (pain) and then pass to knowing about the deeper causal level (how dangerous Kira can get, and how to calm her down). Of course, the action end of immediate self-knowledge (awareness of intention, and then also of desire) is crucially connected to experiencing oneself as a (potential) cause.

When a more complex situation is affecting me, causal connections to myself are much more complicated than in elementary cases, but it is again the causal structure that counts, and I need a reliable model of these causal structures, with self-knowledge as an important focus. Again, I am reacting on the basis of expectations of causally organized course(s) of events, this time starting from me. Note the structural similarity and continuity between the deeper CD-level (with the knowledge of it) and the immediately accessible occurent states (and knowledge of them). Causal connections are much more complicated than in elementary case, but it is again causal structure that counts. Similarly, with contingency planning: reacting in thought to imagined, possible situations, the imagination of possible situation is affecting me. And for this I need a modally rich and flexible view of myself, which is exactly what developed CD self-knowledge is supposed to offer. Again, I am reacting on the basis of expectations of causally organized course(s) of events affecting me or starting from me. Integrated CD self-knowledge the final product would be integrated knowledge of

one's cd-structure. (J. Campbell was on the right track, but did not connect elementary and basic processes to the issues pertaining to more substantial self-knowledge).

Let me just mention where CD-self-knowledge would fit into the well-known scheme due to John Perry. He distinguishes three kinds of self-knowledge (Perry 1998: 83):

(a) agent-relative knowledge,
(b) self-attached knowledge and
(c) knowledge of the person one happens to be.

(a) Agent-relative knowledge is "knowledge from the perspective of a particular agent." It does not involve a term referring to the speaker-thinker, so a sentence like "There is an apple" would do as an example. However, "to have this sort of knowledge, the agent need not have an idea of self," so the CD-knowledge certainly would not belong here. In (c) knowledge of the person one happens to be, the agent is represented to herself in just the same way that other people are represented to her. So, (c) is not a good niche for CD-self-knowledge. This leaves (b): self-attached knowledge, where "the agent has an idea of self," which is associated with what Perry calls a self-notion, namely one associated with the identity of the speaker-thinker. This seems the right location for CD-self-knowledge.

Let me conclude this descriptive-classificatory section with a brief reminder of various sub-kinds of self-knowledge, most of which will not be much discussed here, in addition to those mentioned at the beginning. We had mentioned just two, quoting Perry, agent-relative knowledge and knowledge of the person one happens to be. The first is very much like the phenomenal one, only it concerns one's propositional attitudes: "I know that I believe that three plus two equal five." A related kind of self-knowledge concerns knowledge of (self)-identity through time: the amnesiac Rudolf Lingens (from Perry-Frege's famous example, originally in Frege's "Thought"), can locate himself, say in the library, can know that he feels warm or nervous, and that he means water by "Wasser," without knowing who he is, in the sense of knowing basic data about his identity through time, i.e. his life-history, maybe his name, and so on. He lacks a particular sub-kind of what Perry would describe as self-attached knowledge. Finally, there is the third person philosophical-cum-scientific knowledge of what our self really amounts to, the knowledge of "self itself" to use a phrase from *Alcibiades* 129a, transposing it into the present-day context (think of Damasio's two books on the self (see References), or Galen Strawson's (2011) work on the topic). Each of these raises the interesting issue of epistemic value, and I hope to address them some time in the future. We now return to our two chosen kinds of self-knowledge and pass from factual to evaluative-normative considerations.

2.2. Value and Self-Inquisitiveness

We now pass to value. Let us assume that the epistemic value of a piece of knowledge can be extrinsic, mostly instrumental, or intrinsic, namely the value that the piece has in itself. We have been talking about two kinds of self-knowledge. Combining this contrast between phenomenal and CD-knowledge with the one between two kind of epistemic value, intrinsic and extrinsic, we hit upon four possible combinations (and corresponding characterizations) that will be relevant here: first, extrinsic (practical, instrumental) value of knowledge of phenomenal occurrent episodes; second, their intrinsic value; third, the extrinsic value of CD-self-knowledge; and fourth, the intrinsic value of CD-self-knowledge. Are all these possibilities actualized, and how?

Let us start with direct inner-state self-knowledge. What is the value of such self-knowledge, in particular of its phenomenal core? I shall address it briefly, with apologies for brevity. Cassam, who dedicates a whole chapter (Ch.15) of his (2014) book to the value of self-knowledge, does not say anything about the positive value of direct inner-state knowledge, which he classifies as a kind of "trivial knowledge." His main point of discussing it at all (in Ch. 4) is to show that it is too bland to be philosophically interesting. However, it is pretty clear that it does have extrinsic practical value. Just look at its phenomenal sub-kind: humans react to experienced pain, thirst and the like. Congenital insensitivity to pain is, understandably, described as an extremely dangerous condition. Thus, the extrinsic or instrumental value of phenomenal self-knowledge goes from the most elementary but also most important biological goal, survival, to the general hedonic value of procuring enjoyment, higher hedonic value of aesthetic enjoyment and so on.

Someone might object that qualia are epiphenomenal, and thus cannot procure all the goods we are ascribing to them. Hopefully, our epiphenomenalist will have an account of what distinguishes the behavior of a person with congenital analgesia from the behavior of the normal one. Presumably, she will talk about the pain-realizer states as having a causal role, and claim that such states are absent in the analgesia case. If this metaphysical solution is the right one, the value just moves to a metaphysically lower level, and we might talk about the value of having the realizer states.

So the practical instrumental value, including survival value, is quite unproblematic. What about the other kind, the intrinsic one? Well, start with analgesia, and imagine the area of immediate, phenomenal sensitivity shrinking further: you stop being aware of hearing anything, of feeling [?], and so on. The inner light is being replaced by the "darkness within," to use the famous expression due to McDowell (1998: 250). It is "impossible

not to be concerned with it," he says in a slightly different context (1998: 251). If the phenomenal light within were replaced by such a darkness, you would turn into a zombie, and stop being who you are. So it seems that this kind of self-knowledge has intrinsic value in a very strong sense of "intrinsic': *it is internal to our self through being constitutive of it.*

Here is a possible objection. Coliva, in her (2016) book, in Chapter six on "Constitutive Theories," mentions that "all constitutive theorists agree that the following (scheme of a thesis) is an *a priori* conceptual truth.

> Constitutive Thesis: Given C, one believes/desires/intends that P/to φ iff one believes (or judges) that one believes/desires/intends that P/to φ Coliva, (2016: 170)
> C-conditions must be characterized by reference to subjects who possess normal intelligence, rationality and are endowed with the relevant psychological concepts. (Coliva 2016: 22)

Now a constitutive theorist might want to turn the tables here, and argue that there is no specific value to self-knowledge here, since the knowledge state is not distinct from the known item, and thus does not add anything to it: knowing one is in pain is just being in pain. She might simply proclaim that:

> Given C, one is aware of being in the state of pain iff one is in pain.

And the right side explains the left side completely; there is no surplus fact, and therefore no surplus value to the left side. Coliva, for example writes:

> … very few theorists would subscribe to the view that we constitute sensations and perceptions or simple basic emotions by judging that we are enjoying them, … (Coliva 2016: 170)

However, we are not claiming that judging does the job; it is rather the awareness, a primitive form of self-knowledge that does it, and it is indeed constitutive of the right-hand side: if you are not aware that you are in pain, then you are not in pain, period. Even if we are wrong and it is judging that does the job, the constitutional account does not necessarily devalue self-knowledge. Some "constitutionalists" see the constitutive role as part of the importance and value of self-knowledge. Part of "what we respect in respecting persons, is their capacity for self-knowledge," writes Charles Siewert (2003: 145) in conclusion of his defense of a version of constitutive argument (focused upon higher cognitive states, such as beliefs and desires).

To sum up, with apologies for brevity, the immediate phenomenal self-knowledge has unproblematic practical value, going from the importance of survival to everyday practical concerns. It also has a high intrinsic value since it is constitutive for our self's being what it is.

We now pass to CD self-knowledge, of course in its self-attached variety, of the kind illustrated by "I am very sensitive to dental pain," or "I am prone to jealousy." First, its practical extrinsic (instrumental) value is quite clear. The causal circle of being affected and reacting properly is crucial for survival, and at least some implicit knowledge of it is essential for control and for flexibility. So this might be the biological point of simple CD self-attached knowledge. Even non-human animals might enjoy it in some primitive form, tied to simple bodily self-awareness. (Bermúdez 1998: Ch. 6)

For humans, it is often the case that a complex situation is affecting one. Here, causal connections to oneself are much more complicated than in elementary cases, but it is again causal structure that counts and the agent needs a reliable model of these causal structures, with self-knowledge as an important focus. Again, one is reacting on the basis of expectations of causally organized course(s) of events, this time starting from oneself; time for contingency planning, reacting in thought to imagined, possible situations. The deeper the agent goes in self-locating and self-concerning thought, the richer and more interesting the causal structure gets.

As before, the action end of immediate self-knowledge (awareness of intention, and then also of desire) is crucially connected to experiencing oneself as a (potential) cause. Causal connections are much more complicated than in elementary cases, but it is again causal structure that counts. Here I, the agent, need a modally rich and flexible view of myself, which is exactly what developed self CD-knowledge is supposed to offer. And again, I am reacting on the basis of expectations of causally organized course(s) of events affecting me or starting from me. Central to the CD-level is the causal structure of one's self. One causes things, acting in the world, and also, crucially important for wisdom and care for the self, acting on oneself. So the CD-level concerns causally oriented active and passive dispositions of one's self, as well as the ways of being (possibly) affected by various courses of things, and of reacting to them. Agents have at least three kinds of reasons to look at self CD-knowledge: first, a wide range of practical applications, from survival to small needs and pleasures. Second, the practical importance of a "self-critical perspective," crucial for ethics and views of human happiness and welfare. The point of it all might be the care of the self, finding and realizing the most meaningful kind of life for myself. CD-self-knowledge obviously has a wide range of practical application, from survival to wisdom. (On the theoretical side we have at least two kinds of motivation: first, the traditional philosophical interest: making sense of classical views of self-knowledge; and second, the interest of psychology as science.)

Note that in normal cases one's self-knowledge interacts with one's knowledge of the social surroundings, since the causal chains very often extend in this direction. The practical value of such a wider knowledge is crucial for our practical interests.[7]

However, there is more. Consider cases of complete lack of self-inquisitiveness. Take an up-to-date example: Jane is a nice but totally unreflective person, often volunteering to help refugees. However, she is totally uninterested in the sources of her own motivation. Moreover, she is incapable of giving even a minimally general account of the reasons why she did or omitted some action. "Jane, why did you volunteer?" "God knows, I just felt like doing it!" And this is it. Once we consider the negative value of such self-blindness and self-disinterest, we can see that CD-self-knowledge does have a value that goes beyond its practical consequences, a kind of intrinsic value. Its having intrinsic worth becomes obvious. Substantial, CD-self-knowledge does have some intrinsic value.[8]

It seems thus that all four combinations we started with are actual: there is enough epistemic value around for our two kinds of self-knowledge

3. Discussion: More on Self-Inquisitiveness and the Value of CD-knowledge

Of course, the brief suggestion from the conclusion of the preceding section is just the beginning of the debate. Philosophers have been recently asking a lot of question about the issue. They have concentrated on a slightly wider question: does self-knowledge of the CD variety have any value higher than an immediate practical one? Does it have some higher

---

[7] If you need a quick reminder of how dramatic the interaction of the subjective and social can get in one's life, a good place to look at is *The Autobiography* of Malcolm X; there is some material stemming from reflection, but it is dramatically interwoven with awareness of extremely negative external circumstances in which the author has been growing.

[8] Eric Schwitzgebel has been offering a similar argument from the first-person perspective *Argument 1: The Argument from Addition and Subtraction.* "…subtract: Right now I think I know about myself that I'm kind of a middling extravert and a kind of a middling racial egalitarian with, probably, an ordinary middle-class-white-guy set of implicit racial biases. Subtract this knowledge. I have no idea whether I'm an introvert or an extravert, or I wrongly think I'm an introvert. Stipulate again: no practical consequences. Or suppose I have no idea where I am in implicit racial egalitarianism; maybe I falsely think I'm wholly bias free. Suppose again, no practical consequences. Isn't something important lost?" Eric Schwitzgebel "The Intrinsic Value of Self-Knowledge" February 6, 2015, available on author's web page. And he concludes very affirmatively: "self-knowledge, when we have it, is one of the most intrinsically valuable things in human life." (Ibid.)

extrinsic value? And does it have any intrinsic value? I have been claiming that it has both kinds. Now it is time to address the arguments recently offered against any kind of non-immediately practical instrumental value of knowledge of one's long-term dispositions, by authors like Feldman, Hazlett and, above all Cassam.

Before proceeding to the topic, let me note that these criticism have recently found a parallel in the semantics and epistemology of self-reference, in the work of Herman Cappelen and Josh Dever (2013), with a telling polemical title: *The Inessential Indexical: On the Philosophical Insignificance of Perspective and the First Person*. They describe the goal of their book as consisting in showing "that the entire topic is an illusion—there's nothing there." (2013: 3) They note that (...) indexicals appear to be devices that put us in a uniquely direct and primitive kind of contact with ourselves ("I"), with important features of our environment (the time and place we are at, as with "now" and "here"), and with objects we can demonstrate ("that"). (2013: 10) But this special status of indexical thought, including *de se* attitudes, and the special perspective connected to self-knowledge is an illusion. The truth is, they claim that "our view on the world is not primarily a view from a perspective. Our beliefs and desires are not organized around us. They are instead organized around the world itself (…). Our view is a view from everywhere." (2013: 180)[9] So much for the parallel attack. We shall stay with doubts about any kind of non-immediately practical instrumental value of self-knowledge, including its intrinsic value. We shall see that doubts about "high-style" extrinsic goals for self-knowledge, such as authenticity, go together with doubts about the intrinsic value of self-knowledge in general.

Let me start with a short story taken from Simon D. Feldman and Allan Hazlett (2013), also use by Cassam in his criticism of any higher value of self-knowledge. In the story, Sam "is stuck in a dead-end philosophy job in Boringtown, Connecticut. He has recently had a passionate affair with Grace, a visiting speaker from the exotic University of the Mediterranean. Grace has returned home and it's unclear whether they'll ever see each other again. But Sam doesn't want the romance to end. He is tempted to skip town and join Grace at her seaside villa, but knows that this would be the last straw with the tenure committee at Boringtown State College, given his lackluster teaching evaluations and non-existent publication record."

---

[9] And they continue: Our nature is not deeply as vantages on the world, but as one among many occupants of the world. We see ourselves along with everything else. Some things are seen more clearly and some less. The unclarities in our views of ourselves are at least as prominent as the clarities. So then, what of perspective and indexicality? These are real phenomena, they just aren't deep phenomena. (Ibid.)

(2013: 177) Sam "reflects and introspects, trying to figure out what he cares about: Grace, his career" or something else. After a lot of reflection, he concludes "I am in love with Grace, therefore I shall go on a tryst," and heads for the airport. We are next invited to compare his counterpart, "unselfconscious Sam," also in love with Grace: his story is the same, we are told, "minus the self-investigation and minus the self-knowledge," but with the same resulting action. Sam makes his decision spontaneously, not based on his self-knowledge. Feldman and Hazlett conclude:

> We submit that unselfconscious Sam enjoys a species of intuitively appealing authenticity, which self-conscious Sam lacks. The difference comes down (at least in part) to self-knowledge: unselfconscious Sam lacks self-knowledge, while self-conscious Sam has self-knowledge, and acts on its basis. To put this another way, self-conscious Sam suffers from having "one thought too many (Bernard Williams)." (Feldman and Hazlett 2013: 177)

Feldman and Hazlett use the story to defend the view that self-knowledge has no deeper value. Quassim Cassam joins them in his (2014) *Self-Knowledge for Humans*, in particular Ch. 15. "The Value of Self-Knowledge."

I want to defend the opposite line: "substantial" self-knowledge does have both a high-level instrumental and some intrinsic value. First of all, I must admit that I don't find self-conscious Sam, as described, lacking in authenticity. "I love her, so I shall do what she asks me to do"; if having such a thought means being inauthentic, most of us are very inauthentic indeed. The existentialist thinkers, notorious for being obsessed with authenticity, assume that their heroes know a lot of relevant general truths about themselves and their lives, and recommend having such knowledge. Here is Camus, praising the hero who realizes that his (sic!) life is absurd:

> Living an experience, a particular fate, is accepting it fully. Now, no one will live this fate, knowing it to be absurd, unless he does everything to keep before him that absurd brought to light by consciousness. Negating one of the terms of the opposition on which he lives amounts to escaping it. To abolish conscious revolt is to elude the problem. The theme of permanent revolution is thus carried into individual experience. Living is keeping the absurd alive. Keeping it alive is, above all, contemplating it. (Camus 1955: 38)[10]

---

[10] And here is more: "If I convince myself that this life has no other aspect than that of the absurd, if I feel that its whole equilibrium depends on that perpetual opposition between my conscious revolt and the darkness in which it struggles, if I admit that my freedom has no meaning except in relation to its limited fate, then I must say that what counts is not the best living but the most living. It is not up to me to wonder if this is vulgar or revolting, elegant or deplorable. Once and for all, value judgments are discarded here in favor of factual judgments. I have merely to draw the conclusions from what I can see and to risk nothing that is hypothetical. Supposing that living in this way were not honorable, then true propriety would command me to be dishonorable." (40)

But let us grant for the sake of argument that self-conscious Sam is less authentic. It seems to me that the only way to understand him this way is to assume that he uses his self-knowledge as a *means to reach the decision*, and that it is the wrong means in the situation. He should have been prompted by his love, not by his reflective, second-order knowledge that he is in love. However, in spite of the link with authenticity, this *concerns the instrumental role of self-knowledge* more than its intrinsic value. Cassam is aware of this:

> The idea that substantial self-knowledge is valuable because it promotes well-being isn't the only way of making sense of the notion that its value is extrinsic. You can imagine a high-minded philosopher who believes that the true value of self-knowledge derives from its links with "higher" ideals like authenticity and unity. To be authentic is to be true to yourself, and the suggestion might be that you can't be *true* to yourself unless you *know* yourself. (Cassam 2014: 211)

And Cassam is keen to uphold the thesis that self-knowledge does have instrumental value. So his only way out would be to claim that the way self-conscious Sam is using his self-knowledge is wrong, not that self-knowledge is generally instrumentally worthless.

There is a related local difficulty connected with authenticity and related virtues. Take modesty. If you believe you are truly modest, you are probably not modest at all. Some virtues thus appear as possible blindspots for self-knowledge. This is probably the deep response, evolutionary and/or social, to the ubiquitous threat of narcissism. However, the problem is local, and just shows that some local self-ignorance is better that self-knowledge on the same local topic. Another example, due to my colleague and friend Danilo Šuster, concerns constitutive impossibility: I shall do this spontaneously, since I am such a spontaneous person! However, the blindspots are not peculiar to self-knowledge. Life is full of similar object-knowledge related, not self-knowledge related analogues, as well as more general cases of demotivating, even paralyzing knowledge of various truths. Hazlett did use them to question the value of knowledge in general in his (2013) book. But *pace* Hazlett, they do not call the value of knowledge into question; the same should hold for local self-knowledge blindspots. So much for Feldman's and Hazlett's example. Let us pass to the next kind of value: intrinsic value.

Cassam argues against any kind of "high-road" approach (although the ones he mentions are all on the extrinsic side) by arguing against the claim that self-knowledge is needed for the authenticity and unity of one's life. And indeed, if I need self-knowledge in order to be authentic, then the resulting value looks like instrumental value, not intrinsic, as we just saw in

the Sam and Grace example. Remember, he is very much in favor of "low-road" practical value: self-knowledge helps us live a good life. Bad luck: it has no intrinsic, non-practical value.

Let us then consider this possibility in more detail. In general, the anti-value arguments sound good for isolated examples. In novels, examples of first-order wisdom (moral correctness, authenticity) unaccompanied by reflection are relatively "exotic": prince Myshkin, from Dostoyevsky's *Idiot*, and Platon Karataev from Tolstoy's *War and Peace*, nicely illustrate the purely first-order governed life. The first is too ill, and the second not sufficiently literate to develop a theory-like second-order wisdom, but both are practically wise in a basic and very attractive way, spontaneously and correctly doing and advising what is right in the most dramatic situations of their lives. Here is Tolstoy's characterization of Karataev:

> Each of his words and each of his acts was the manifestation of an activity he knew nothing about, which was his life. But his life, as he looked at it, had no meaning as a separate life. It had meaning only as a part of the whole, which he constantly sensed. His words and acts poured out of him as evenly, necessarily, and immediately as fragrance comes from a flower. He was unable to understand either the value or the meaning of a word or act taken separately. However, the critic of intrinsic value of (CD) self-knowledge faces two problems. First, is not our admiration simply the result of the relative strength of two kinds of values–isn't it the case that Myshkin and Karataev have first-order moral qualities so admirable that they outweigh the lack of reflection? Imagine a continuation of *War and Peace* in which Karataev joins Tolstoy's commune and become a wise thinker, retaining his kindness and his capacity for spontaneous moral and prudential insights. Would this involve a loss of value? I doubt it. Some gain? Probably. (Tolstoy, *War and Peace*, Vol. 4: 282)[11]

The second problem is equally worrying for our critic of the intrinsic value of substantial self-knowledge. We admire Myshkin and Karataev and enjoy reading about them. But is a relatively wide, blissful, reflective 2nd level ignorance possible for us, reflective creatures? The two Russian heroes mentioned are quite unlike us. We want to know about ourselves. Indeed, the Karataev admirer in *War and Peace*, Pierre Bezuhov, is in this respect one of us: he learns from Karataev, but he systematizes what he has learned on the second, reflective level. (Wittgenstein was apparently another admirer of Karataev, and wrote philosophical, reflective, and sophisticated comments about the beauty of being unselfconscious.)

---

[11] Karataev's Russian admirers go as far as to claim that the un-selfconscious nature of his actions brings them close to non-acting, as Wladimir Kantor puts it. (2010:130 ff) ("Das gilt auch für Karataev. Sein Tun ist nicht bewusst und steht daher dem Nichttun nahe ..." to quote the German text.)

So how unselfconscious can one be without becoming shockingly blind to oneself? Remember Jane, think of amazing Grace, and look at truly unselfconscious Sam. You ask him:

- Sam, why did you go to Greece?
- Well, in fact, I don't know. I just felt this pull, which had something to do with that woman, Grace
You persist:
- Did you fall in love with her?
And he answers:
- Hard to say, I never know why I do things.

Starts sounding a bit problematic, doesn't it? Like a case of complete apathy about coming to know about oneself, epistemic self-apathy if we may so call it. Maybe it is perfect for a hero from Dostoyevsky or Tolstoy, but it would certainly be far from perfect for an average university student. Wittgenstein found the inability to reflect impressive in the case of Karataev; he was much tougher with his Cambridge colleagues and pupils, whom he pushed incessantly to reflect and to justify their views. In short, we simply expect a degree of self-reflectiveness from people from our surrounding. It is normal to have a degree of non-practical interest in one's long-standing motives, in one's character, and in other causal-dispositional properties we are focusing upon.

Why do we feel that Jane and truly unselfconscious Sam miss something important? One answer is the expectation of coherence. A person should have a coherent view of oneself, a "loop" (Lehrer 1997) that will hold all elements together, theoretical and practical alike, and exercises some control over the first level. So the mutual support of knowledge and motivation, the theoretical and the practical, points to a second-level fully reflective system, as does the need to balance the full range of contrasting considerations, prudential, moral, and meaningful life-related.

Otherwise one ends up like unselfconscious Sam, incapable of asking about the coherence of one's wishes. Here is one more possible continuation of the conversation with him:

- Sam, do you really never try to understand why you did things, even very important ones, like joining Grace in Greece?
- No, why would I?

Something is missing. Even worse, something is amiss. One element that is missing is coherence. In order to live wisely, one has to fulfil a first-level and a second-level condition: on the first level to have correct action-guiding

preferences, and on the second level a coherent reflective mechanism that balances moral, prudential and meaningful life-related considerations, can take control if needed, and gives the agent a coherent reflective perspective of oneself and one's situation. We simply do normally enjoy some amount of self-curiosity: self-insight is cherished by people, no matter how difficult it is to achieve.

Now what would psychologists say? Here are a few concluding sentences of Dunning's book on self-insight:

> Life presents many challenges, strewn like hills and mountains in our path. Acquiring self-insight might just be one of those peaks, one that is more rugged and steep than it looks from afar. ...From its peak, one does not know what the view of the psychological terrain might look like, but I assume that most people would be quite curious to take a glimpse of this view. (Dunning 2005: 184)

Dunning stresses the role of interest and curiosity, as one would expect when it comes to intrinsic epistemic value. Let me translate this talk into the usual epistemological vocabulary. Epistemologists talk about epistemic virtues in two senses: some describe crucial cognitive abilities as virtues, others concentrate on epistemically (and often morally) positive character traits. Self-knowledge needs both kinds. On the one hand, we need ability-virtues to reach it; on the other we need positive curiosity. Let me call it "self-inquisitiveness." Epistemic apathy or sloth about oneself would be the opposite of it. Self-inquisitiveness is an epistemic virtue of the character-related type, tied to the value of self-knowledge. Note that CD-self-knowledge (including knowledge of one's merely dispositional beliefs and desires) seems the most natural object of self-inquisitiveness. Self-locating knowledge might be another worthy target in rare, and less natural, problematic situations, such as the Sleeping Beauty scenario. Direct second-order knowledge of one's phenomenal states and occurrent beliefs and desires needs no special effort

So let me conclude this section with a question that naturally arises at this point. How do people slide from extrinsic interests to the corresponding intrinsic ones? Assuming that CD-self-knowledge has practical value, where would its intrinsic value come from? Why do we consider self-blind people, victims of epistemic self-apathy, such as Jane and truly unselfconscious Sam, to be missing something essential? Psychologists talk about the functional autonomy of motives: a motive that started as instrumental-extrinsic can take over and become a goal in itself. Our curiosity about external matters probably proceeds in this way. People were curious about the details of the starry heavens for purposes of maritime travel and the like; at some point a "pure" interest in astronomy developed. Interest in

one's habits and character might have bifurcated in a similar fashion: the older, purely practical interest remained, but a new kind of curiosity, an intrinsic one, developed. We can apply it to self-knowledge as well. CD-self-knowledge is essential for practical purposes, including short-term and long-term planning. We can imagine that this extrinsic importance has made people intensely curious about their habits, character and the like. The curiosity paid off; the better they knew themselves (in this "low-road," modest way), the more successful they were at large. Once the motive was there, it could have partly detached itself from the original practical framework. People became "curious to take a glimpse of" their psychological landscape for the sake of it.

Notice that the social division of relevant epistemic labor has probably proceeded in a similar fashion. Once you have the possibility of specializing and following your intrinsic motivation, the option of pure research is born. It is usually coordinated with application and embedded in applied frameworks, but in good cases the researcher does not have to worry about these further matters. The same might be valid for study of the self. First, some modest specialization was born, with specialists (priests, healers) taking care of dramatic shortcomings having to do with the understanding of one's motives, drives, and abilities, all for immediate practical purposes. Next, and much later, some non-practical interest in the same topic found a social niche in which to survive; it might have started with priests and poets, but we philosophers, know for certain that at some time it appeared in philosophy. All this is, of course, hypothetical, but it shows that there need not have been any mystery about the birth of intrinsic self-inquisitiveness.

4. Conclusion: The Virtue Epistemology of Self-Knowledge

Let me summarize the main points and then briefly address questions that the reader might have but that have been left unanswered so far. Two kinds of self-knowledge, out of many, have been of interest to us here, immediate phenomenal and CD-knowledge. We can now characterize them in terms of a virtue-epistemological framework. We see the two kinds as valuable, both practically and intrinsically. Their epistemic value suggests that the typical capacities that procure them for knowing subjects, such as self-awareness; sensitivity to one's causal powers, active and passive; and the like, are virtue-abilities. We have also suggested that the values in question are connected to our self-inquisitiveness; in particular, this holds for the intrinsic value of self-knowledge. We have arrived at the following schematic picture of the situation:

		KINDS OF VALUE	
		INTRINSIC	**EXTRINSIC**
OBJECT OF KNOWLEDGE	**Pheno-menal**	• Constitutive	• Practical-survival
	CD	• Self-inquisitive-ness • Prevention of self-blindness	• Low-practical: surviv-al, everyday needs • High: authenticity, wisdom

Let us look very briefly at each box and each combination. First, phenomenal inner-states knowledge. When we talk about its value we do not assume a complete, perfect and luminous knowledge. Even if luminosity is lacking, I might know that the red color I am experiencing is saturated, without knowing the exact degree of saturation. The practical value is generally clear: Mary sees red (coming from the traffic light), she is aware that it is red and she stops. Now a critic might insist that the practical value is completely inherited from the first-order state. If the two are inseparable, as the critic assumes, I see no reason for thus insisting: the value belongs to both, no problem.

Why think that such phenomenal knowledge is intrinsically valuable? One reason is its constitutive role: no seeing red without knowing that one sees it. Another is our natural curiosity. Black-and-white Mary is dying of curiosity about what it is like to see red; others might be similarly curious about experiences in sport or dancing, and, having been reading Lewis in my hometown far away from Australia, I have been intensely curious about what it is like to taste Vegemite (and was disappointed once I did taste it).

Self-attached CD knowledge has extrinsic instrumental value of all imaginable kinds. On a low but immensely important level, it enables our survival, and then the fulfilment of our everyday needs and wishes. On the higher level (the "high-road" kind, as Cassam calls it), it secures the coherence of our mental "make-up" and gives us a consistent picture of it. Those who are authentic in their virtues are made more systematically and intelligently authentic if they enjoy a bird's-eye view of their causal powers, active and passive. The traditional ideal of wisdom includes a very high degree of self-knowledge of one's character. This is compatible with the inevitability of a few local blind-spots, connected with virtues such as modesty or generosity. "I know that I am very generous" sounds very, very fishy. A virtue theory should give an account of why some important virtues should not reveal themselves fully to their bearers.

Self-attached CD knowledge also has intrinsic value: the clear sign of this is that people normally despise systematic self-blindness and complete

apathy concerning one's own traits. Interestingly, the attitude is normally stronger against the lack of self-inquisitiveness than against the lack of objectual acquisitiveness. Take Peter, who completely lacks interest in politics, in art and classical music, and in sports, and compare him to self-blind Jane. Peter is ordinary; Jane is almost pathological. Philosophers who have written confessions, from St. Augustine to J.J. Rousseau and J.S. Mill, have brought testimony to their self-inquisitiveness, their interest in self-awareness exactly of (what we characterized as CD). So much for our two paradigmatic kinds of self-knowledge.

We now pass to two final open questions, to be properly addressed at some future occasion. For example, what about other kinds of self-knowledge? How much intellectual excitement do they promise? Allow me a brief look forward at those kinds of self-knowledge, which I hope to discuss on some other occasion. Consider the kind closest to phenomenal self-knowledge: knowledge of one's propositional attitudes. A constitutivist might argue that there is not much to be added to the value of the first-order attitude: my knowledge that I believe that *p* inherits its main properties from my belief that *p*, and does not add anything interesting. She might remind us of the ascent/descent recipe due to Gareth Evans *Varieties of Reference* (1982) in a chapter devoted to self-identification, and developed by authors like Moran (2001) and Fernández (2013). If I want to know whether I believe that it is raining, I should descend to the first level, and just ask myself: Is it raining? If I answer Yes, I can then ascend to the meta-level and assert: Yes, I believe it is raining. The idea is put forward in *Varieties*: "In making a selfascription of belief, he writes, one's eyes are, so to speak, or occasionally literally directed outward – upon the world." (Evans 1982: 225)

Assume that Evans is right, and that this does speak in favor of constitutivism. But are the evaluative consequences so bleak? Here is a challenging line of thought. One might come to discover, by using Evans's recipe, that one hold a belief one is surprised to hold. Consider the following story, quite a realistic one, I submit. Jim is not a very reflective person. He is a moderate leftist, but he has had unpleasant conflict with two female colleagues at his job, and his bad experience has started coloring his general attitude toward female colleagues. He comes out with unpleasant comments, and at some point his leftist friend asks him explicitly: Jim, do you really believe that women are incapable of performing serious work and of taking higher administrative responsibility? Jim takes it seriously, and asks himself the first-level question: wait, are women really incapable of performing serious work and of taking higher administrative responsibili-

ty? And to his surprise, only thoughts in favor of the positive answer flood into his mind. Jim, a good leftist, is shocked: do I really believe this crap? I should stop this, my goodness![12]

Clearly, Jim has learned something about himself, and the information has a positive epistemic value. *The Evans-type strategy can procure valuable items of self-knowledge*, certainly for dispositional, quietly present beliefs. So it seems that the constitutivist's reservations hold only for the other kind, for explicitly, manifestly held beliefs. But if self-awareness (second-order knowledge) is part of manifestness, and the manifest belief is valuable, isn't second-order knowledge the one that adds the surplus epistemic value? If dispositional, dormant belief does not have the relevant value, but value can be added to it by the procedure of descent/ascent, isn't this procedure itself the awakening of self-knowledge, the value-bestowing act? This is the line I would like to explore about attitudinal self-knowledge.

Finally, just a few more words about the nature of intrinsic epistemic value. We have argued for the intrinsic value of self-knowledge from ordinary expectations about people being curious about themselves, or, more dramatically, failing to be curious. The link between curiosity or self-inquisitiveness and epistemic value is quite strong, I think. It makes itself felt in our appreciation of all kinds of knowledge. I am mentioning it because I haven't said anything about the source of epistemic value, and the knowledge-curiosity link is the right place to look for it. Where does the curiosity-intrinsic value link come from? There are two main options, depending on the order of determination:

> First, the grasping of *p*, (say, knowing, or coming close to knowing) is intrinsically valuable because a person (with the right characteristics) would be curious whether *p*. (response-dependentist account)

> Second, persons are justifiably curious whether *p* because the grasping of *p*, (say, knowing, or coming close to knowing) is intrinsically valuable. (strongly objectivist account)

The usual feeling is that some states of knowing concerning some states of affairs are intrinsically epistemically valuable, and people, if intelligent, well-informed-educated and sensitive, are curious about these states of affairs.[13]

---

[12] Compare the quite different story about Ralph the sexist from Schwitzgebel's "Self-ignorance."

[13] A strong defense of such objectivism about epistemic value can be found in Brady 2009.

But why would such a fact have intrinsic value dictating epistemic axiology? Just postulating that it has one leaves epistemic value unexplained. So one could choose the response-dependent option: curiosity bestows value. But unfortunately some people are intrinsically curious about worthless matters, and it is hard to specify what kind of people would be eligible as judges without falling into the trap of question begging. We don't have to answer the question here; I only point out that answers exist. I would prefer the response-dependentist alternative (see Miščević 2016), but this is a topic for another occasion.

I hope the discussion of the two paradigmatic cases has given some grounds for optimism about the value of self-knowledge and the positive role of self-inquisitiveness. So to summarize it very briefly, the injunction "Know thyself!" is still good advice after two and a half thousand years.

REFERENCES

Ahbel-Rappe, S. & Kamtekar, R. (eds.) (2000). *Blackwell Companion to Socrates,* Blackwel.

Bermúdez, J. L. (1998). *The Paradox of Self-Consciousness*. MIT Press.

Bermúdez, J. L. (2011). "Bodily Awareness and Self-Consciousness." In Shaun Gallagher (ed.) *The Oxford Handbook of the Self.* Oxford University Press.

Brady, M. (2009). "Curiosity and the value of truth." In Haddock, A., Millar, A. and Pritchard, D. (eds.) *Epistemic Value*. Oxford University Press.

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

Camus, A. (1955). *The Myth Of Sisyphus And Other Essays.* Translated from the French by Justin O'Brien. Alfred A. Knopf Inc.

Cappelen, H. & Dever, J. (2013). *The Inessential Indexical On the Philosophical Insignificance of Perspective and the First Person*. Oxford University Press.

Cassam, Q. (2014). *Self-Knowledge for Humans*. Oxford University Press.

Coliva, A. (2012). *The Self and Self-Knowledge*. Oxford University Press.

Coliva, A. (2016). *The Varieties of Self-Knowledge*. Palgrave MacMillan.

Damasio, A. (2010). *Self Comes to Mind: Constructing the Conscious Brain*. New York: Pantheon.

Damasio, A. (2000). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness.* Harcourt: Houghton Mifflin, Mariner Books.

Dell, K. (2001). "Wisdom Literature." In L. G. Perdue (ed.) *The Blackwell Companion to the Hebrew Bible*. Blackwell: 418-431.

Dell, K. (2011). "Proverbs." For *Oxford Encyclopaedia of the Bible*, 2 vols, ed. M. Coogan, Oxford University Press: 183-192.

Dunning, D. (2005). *Self-Insight: Roadblocks and Detours on the Path to Knowing Thyself* Psychology Press.

Feldman, S. D. and Hazlett, A. (2013). "Authenticity and Self-Knowledge," *Dialectica* Vol. 67, Issue 2: 157–181.

Fernández, J. (2013). *Transparent Minds: A Study of Self-Knowledge*. Oxford University Press.

Gertler, B. (2016). "Review of Quassim Cassam's (2014) *Self-Knowledge for Humans*." *Mind* 125 (497): 269-280.

Hatzimoysis, A. (2011). "Introduction," to Hatzimoysis, A. (ed.) *Self-knowledge*. Oxford University Press.

Hazlett, A. (2013). *A Luxury of the Understanding: On the Value of True Belief*. Oxford University Press.

Kantor, V. (2010). *Westlertum und Russland*. ibidem-Verlag.

Kraut, R. (2006). "The Examined Life." In Sara Ahbel-Rappe and Rachana Kamtekar (eds.) *Blackwell Companion to Socrates*, Blackwell.

Lehrer, K. (1997). *Self-Trust*. Clarendon Press.

Lewis, D. (1979). "Attitudes *de dicto* and *de se*." *Philosophical Review* 88: 513–543.

McDowell, J. (1998). "Singular Thought and the Boundaries of Inner Space." In *Meaning, Knowledge and Reality* Harvard University Press: 228-259.

Miščević, N. (2007). "Virtue-Based Epistemology and the Centrality of Truth (Towards a Strong Virtue-Epistemology)." *Acta Analytica* Vol. 22, Issue 3: 239–266.

Miščević, N. (2012a). "Learning about wisdom from Lehrer." *Philosophical Studies* 161 (1): 59-68.

Miščević, N. (2012 b). "Wisdom, Understanding and Knowledge: A Virtue-Theoretic Proposal." *Acta Analytica* 27: 127-144.

Miščević, N. (2016a). "Curiosity – The Basic Epistemic Virtue" in Chienkuo. M., Slote, M. and Sosa, E. (eds.) *Moral and Intellectual Virtues in Western and Chinese Philosophy, The Turn toward Virtue.* Routledge: 145-163.

Miščević, N. (2016b). "Epistemic Value-Curiosity, Knowledge and Response-Dependence." *Croatian Journal of Philosophy* Vol. 15. No. 48.

Moran, R. (2001). *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton University Press.

Perry, J. (1998). "Myself and I" in Stamm, M. (ed.) *Philosophie in Synthetisher Absicht*. Klett-Cotta: 83-103.

Rowe, C. (2011). "Self-Examination." in Donald R. Morrison (ed.) *The Cambridge Companion to Socrates:* 201-214.

Schwitzgebel, E. (2015). "The Intrinsic Value of Self-Knowledge." February 6, on authors web page.

Siewert, C. (2003). "Self-knowledge and rationality." In Gertler & Brie (ed.) *Privileged Access.* Ashgate.

Stalnaker, R. C. (2008). *Our Knowledge Of The Internal World*. Clarendon Press.

Strawson, G. (2011 revised edition). *Selves*. Oxford University Press.

Tolstoy, L. (2008). *War and Peace.* Transl. by Richard Pevear, Larissa Volokhonsky. Knopf Doubleday Publishing.

5. The Self-Ascription Of Conscious Experience

LUCA MALATESTI

1. Introduction

We can have thoughts that are expressible with sentences such as "I am having the experience of a pain in my elbow," "I am having a conscious experience of red." These thoughts involve the self-ascription of conscious experiences of a certain type on the ground of having such mental states. Philosophers investigate what is involved in the understanding of these thoughts to account for special features of our first personal knowledge of conscious experience and their nature.[1]

In this article, I elucidate some aspects of our understanding of these self-ascriptions by focusing on the concepts that concern the type of conscious color experiences that we have and the self. Within a controversial area of investigation, I characterize concepts, minimally and intuitively, as *ways of thinking* about objects, properties, and other entities. I introduce concepts so understood by means of *that*-clauses reporting ascriptions of beliefs or thoughts. I take that concepts are individuated by the conditions that a thinker must satisfy to possess them.[2]

I maintain that the self-ascriptive thoughts that are here at issue involve thinking about the type of color experiences that we undergo by exploiting a way of thinking about colors that is provided, mainly, by certain *observational concepts*. These latter concepts are applied, and thus individuated, in virtue of the capacities to visually discriminate and thus individuate the colors that objects appear to have. On the other hand, the concept of self,

---

[1] Notably Christopher Peacocke has touched upon this issue over the years within his general approach to the formulation of a theory of concepts. See Peacocke 1992. Also, the defenders of the so-called phenomenal concepts reply have engaged in this type of endeavor when responding to several interrelated anti-physicalist arguments concerning conscious experiences. See the anthologies Ludlow, Stoljar, and Nagasawa 2004; Alter and Walter 2007.

[2] For the general form of this account of concepts and the challenges that it should meet, see Peacocke 2009.

which is involved in self-ascriptive thoughts concerning color conscious experiences, involves thinking about its referent and, thus identifying it, as an object that has internal states that derive from the stimulation of sense organs.

The next section sets out some central assumptions concerning what is involved in investigating our understanding of the self-ascriptive thoughts. I use Frank Jackson's knowledge argument to show how the understanding involved in the self-ascriptive thoughts here at issue depends on having conscious experiences of the relevant type. I argue that this argument shows that several philosophical analyses, by failing to be appropriately sensitive to this dependence, fail to offer an appropriate account of these thoughts. In the third section, I argue that the dependence between self-ascriptive thoughts and having the relevant experiences cannot be spelled out in terms of a, quasi-perceptual, direct awareness that would provide discriminatory and thus identificatory information about our conscious experiences. In the fourth section, I argue that instead the ways of thinking about the type of conscious experiences we undergo depends on discriminatory capacities concerning how the world looks to us when we have these experiences. In the last section, I show that such a way of thinking about the type of experiences that we have suggests that our way of thinking of the self involves identifying it as an entity that undergoes internal states based on stimulations of certain sense organs.

2. Concepts and their Analysis

We can have thoughts about what grounds our occurring experience as an experience of a certain type. For instance, by seeing a certain color, we might focus on the way in which the color strikes us when we see it, and we can think a thought expressible as "I am having an experience of the type involved in seeing this color" or, shortly, "I am having an experience of red."

Explicating the content of this type of thoughts about conscious experiences requires explicating certain concepts that are relevant for that content. This requirement is compatible with two principal accounts of the relation of the content of a thought with concepts.[3] In one account, let us call it Fregean, the content of a thought is made of concepts. In this case, the requirement above has an immediate plausibility. The other account, usually characterized as Russellian, requires instead that properties and relations make up the content of a thought. In this case as well, it is plausible to require that a subject has a thought with certain content if she possesses

---

[3] See Bermúdez 1998: 52.

the concepts relative to the constituents of that content. However, how, in general, can we explicate thoughts and concepts?

It could be assumed that the contents of thoughts and, thus, of the relevant concepts, should be investigated by means of linguistic analysis. Michael Dummett provided a well-known statement of this type of methodology:

> Thoughts differ from all else that is said to be among the contents of the mind in being wholly communicable: it is of the essence of thought that I can convey to you the very thought that I have, as opposed to being able to tell you merely something about what my thought is like. It is of the essence of thought not merely to be communicable, but to be communicable, without residue, by means of language. In order to understand thought, it is necessary, therefore, to understand the means by which thought is expressed. (Dummett 1978: 442)

There are several proposals in the history of analytic philosophy of mind that are based on this methodology. Their proponents have focused on what is involved in the abilities to describe linguistically our conscious experiences.

Gilbert Ryle offered a paradigmatic and sustained example of the analytic methodology in philosophy of mind focused on how we talk about these states. (Ryle 1949) He argued that talk about conscious mental states concerns, without residue, manifest behaviors and multiple dispositions to behave. Adopting a similar methodology J. J. C. Smart, instead, argued that ordinary ways of talking and thinking about conscious experiences concern types of internal states of the subject that are individuated by their causal mediation of certain types of stimuli and responses. (Smart 1959) Furthermore, functionalists, such as Hilary Putnam, adopted similar analyses or regimentations of mentalistic expressions. (Putnam 1967) However, all these linguistic analyses can be challenged by using a standard analytic procedure. The standard procedure for testing philosophical analyses:

> is to think up a possible general characterization of the cases falling under some concept C and then to test it by trying to find or imagine a particular situation which fits the suggested characterization and yet would not be a situation to which C could be truthfully applied. (Overgaard, Gilbert, and Burwood 2013: 85)

Several philosophical arguments against the analyses of language used in the self-ascription of conscious experiences, which I have briefly described above, rely on this procedure.

The knowledge argument (from now on KA) is a famous objection to physicalism, the thesis that everything is physical, that puts pressure on all the previous linguistic analyses. (Jackson 1982) This argument exploits the case of the scientist Mary. Being confined in a monochromatic environ-

ment, she has never had a color experience. However, the argument goes, it is conceivable that Mary knows all the causal roles, stimuli, and behaviors that are mentioned in the philosophical analyses that I have mentioned above. In addition, she can know all the physical facts concerning the neurological bases of vision. Now, we are not inclined to ascribe to Mary, before she sees colored objects, understanding and knowledge of what is involved in the experience of seeing colors. Thus, the main intuition is that these analyses cannot accommodate the ordinary notion of color experience. As Jackson puts it:

> After Mary sees her first ripe tomato, she will realize how impoverished her conception of the mental life of others has been all along. (Jackson 1986: 292)

If we take the KA as an instance of the procedure for testing philosophical analyses of concepts, it could be maintained that while we are inclined to ascribe to Mary, before her release, concepts concerning color experiences that are analyzable in terms of behaviors, functional roles, or physical states, we cannot truly apply in describing her scientific knowledge the ordinary concepts that we employ in thinking about our experiences when we have them. As Jackson when he proposed the argument, some think that this conceptual difference authorizes the ontological conclusion that Mary's conception of color experiences, before her release, is impoverished because she does not know about certain non-physical properties of her experience, usually indicated as *qualia*. These properties, supposedly, characterize the *phenomenal character* of the experience or *what it is like* to have that mental state.

A different interpretation of the knowledge argument can be based on a different way to test conceptual analyses, which is independent from the referents of the relevant concepts, and thus, the truth of the thoughts where they are implicated:

> **Distinctness of Concepts**. Concepts C and D are distinct if and only if there are two complete propositional contents that differ at most in that one contains C substituted in one or more places for D, and one of which is potentially informative while the other is not. (Peacocke 1992: 2)

The principle above individuates concepts based on their cognitive significance for a subject. Understanding the content of thoughts requires understanding of concepts. It can be maintained that understanding of the content of thoughts presupposes understanding of concepts individuated in terms of their cognitive significance or the *way* in which they allow the thinker to think about their referents. Thus, a procedure for individuating concepts, and thus the contents where they are implicated, is that of find-

ing two complete propositional contents where a substitution of a concept with another would be informative for a subject.

Applying the procedure for distinctness of concepts to the KA delivers a conceptual distinction between concepts expressed by neurological, functional or behavioral descriptions and concepts that enter in the thought we have about the type of experiences that we undergo. Let us call *phenomenal concepts* the latter type of concepts.[4] In fact, while Mary in the room might not find informative the propositional content:

(1) The phenomenal character of the experience of red is property Q (where Q could be a functional role or brain state XYZ or a relation R between the brain and the environment).

It is plausible to maintain that the KA elicits the plausible intuition that, when seeing a colored object for the first time, she would find informative the proposition:

(2) The phenomenal character of the experience of red is this.

In the content expressed by sentence (2), "this" refers to whichever property we ordinarily think about, when we have a certain experience, that grounds our thought about the type of conscious experience that we are having. If this reading of the situation involved in the KA is correct, it seems that the analysis of the content of thoughts concerning the self-ascription of types of conscious experiences should involve an analysis of the thought that is expressed by sentence (2). But, how can we analyze the concepts involved in that thought?

It might be suggested that the analysis should concern a linguistic canonical expression of the demonstrative concept involved in the thought about the experience at issue. Thus, it could be maintained that Mary, before being released, would be capable, independently from having that experience, to have thoughts involving the concept characterized by that description. However, in any case, a plausible intuition is that when Mary would see a colored object, she would find an informative difference between referring to her experience by using the concept expressed by means of that description and referring to it by using the one that requires having the color experience. Therefore, it seems that analyses based on the linguistic description do not explicate the contribution of her demonstrative concept to the content of her self-ascriptive thought. There should be another approach to capture that way of thinking that requires a role for the experience in determining the type of content of the thoughts about the color experience that she is having.

---

[4] This interpretation is central in the so-called *phenomenal concept reply* to the KA. Versions of it can be found in Tye 2000, Papineau 2002, and Perry 2001.

Instead of considering the linguistic canonical expression of the thought involved in thinking "I am having an experience of a certain type," we should consider that the relevant ingredients in our understanding of the concept of a conscious experience should be determined by assuming the following principle:

> **Simple Formulation.** The relevant concept of conscious experience is that unique concept C for the possession of which a thinker must meet the condition that she has had experience e.

Clearly this account needs to explain in more detail how undergoing the experience is a requisite for the possession of that demonstrative concept. In the next section, I consider one account to then dismiss it.

3. Against the Direct Identification of Conscious Experience

It could be maintained that having a certain conscious experience is a necessary requirement for possessing a certain concept because it confers on the agent certain capacities. One account would be that having that experience would confer the capacity to detect directly some identifying information about the referent of that concept.[5] In general, this is information about an entity that is needed to discriminate it from other entities. Consider, for instance, when someone thinks about a certain car with a thought expressible as "that car is old" based on his perception of the car. It seems that the demonstrative way of thinking about the car involves, amongst other things, the car as it is perceptually discriminated from the rest of the perceived scene. In the case of conscious experience thus, an account of how we can come to think:

(1) I am having a conscious experience e,

would assume that the mastery of the concept of conscious experience e involves the capacity to be directly aware of e. This is to be understood as the capacity that involves, amongst other things, discriminating e from other entities that is based on some information made available in the act of awareness. Thus, for example, to think:

(2) I am having the experience of a red table cloth,

we should be directly aware of the experience and apply the concept because of such awareness. Given that perception can be regarded as offering identification information about its objects, this account could be called a

---

[5] The suggestion of this account is inspired by the work of Gareth Evans on demonstratives, see Evans 1982.

quasi-perceptual account of the epistemic relation with the experience, that, in turn, is required for the self-ascriptive thoughts that I am here investigating. For instance, materialists such as David Armstrong have posited the existence of an inner scanner in the brain, where a second order state makes us aware of the experience at issue. (Armstrong 1968)

This account might also be extended to accounts of the mastery of the notion of self that is involved in the self-ascription of experiences.[6] So, the account of the content of the relevant thought would have the following general form:

> (A1) S's thought, expressed or expressible as "I am having an experience of a certain type" requires S's capacity to be directly aware, by means that enable S to use inner demonstratives, of herself and of the type of experience that she is undergoing.

Subjects would be able to think about the fact that they are undergoing a certain type of experience by being directly aware of their self and that experience. Within this picture, it could plausibly be added that the subjects are also directly aware of the fact that their self has that type of experience.

The quasi-perceptual account is flawed; the difficulties with it emerge when we try to specify in more detail the notion of direct awareness that it requires. Following a suggestion by Sidney Shoemaker, a plausible way to delineating the notion of the awareness of conscious experience would be to rely on intuitions concerning the ordinary understanding of the awareness of objects in our perception of the world, understood naively. (Shoemaker 1996) It seems that one of the features that characterize directness in the context of perception is the capacity to formulate demonstrative thoughts about the object at issue. Thus, by extending this aspect of perception to introspection, the direct awareness of conscious experiences would involve the capacity to have demonstrative thoughts about experiences and the features that ground their categorization. We should be able to discriminate the experience, based on certain identifying information, from a certain background of other mental states or experiences. Now, the resources that we have for discriminating our experiences, and thus typifying them, from a background of other experiences, do not appear to relate to "inner" acts of demonstrations directed to those mental states that are related to discriminatory capacities to support a demonstrative way of thinking about them.

---

[6] This is not necessary; an inner sense account of our epistemic relation with conscious experiences might be coupled with different views on our epistemic access to the self.

Inner demonstrative acts of experience cannot make sense of our practices of describing the type of conscious experience that we have, and, plausibly, of reporting the content of our thoughts about them. (Millar 1991) Thinking and saying, for instance, that I am having a conscious color experience would require that I am performing an act of "inner pointing" to that experience, as discriminated from other mental states, to fix the type of experience I am thinking or talking about. However, this would barely inform other people about the type of experience that we are having. Instead, our practices of informing others about the type of experiences we have, by expressing our self-ascriptive thoughts, rely on typifying our experiences in terms of typical situations that mention aspects of the situation we are in when we have those experiences. So, we might say things such as "I am having an experience of red" or "I am having the experience of drinking Chianti." These descriptions appear to inform someone correctly about the type of mental state we have, if she had that type of experience under the appropriate stimuli conditions and she masters the concepts involved in our description of the relevant situation.

In addition, it seems that "introspective" evidence supports the view that we are not directly aware of our experiences. G. E. Moore in a classical passage, where we can substitute for his notion of sensation our notion of experience, observed that:

> When we try to introspect the sensation of blue, all we can see is the blue: the other element is as if it were diaphanous. (Moore 1922: 25)

As some say, the experience, as a perfectly polished window, is "transparent" to the gaze involved in direct awareness. Thus, when thinking about the type of color experience that we are having, based on what is visually available by having that experience, we can only be directly aware, and thus discriminate, certain qualities. However, these do not enable us to discriminate the experience in the same way in which perceived features of an object enable us to discriminate that object. The experience *as* such is not an object which we can directly discriminate in virtue of the properties we are aware of in having the experiences. While we are directly aware of the properties that ground the self-ascription of the type of experience we are having, we are not directly aware of the conscious experiences and of the fact that they have the properties that determine what type of experiences they are. Thus, at least in this respect, the analogy between the awareness of conscious experiences with perception breaks.

I am not engaging here with the issue whether transparency reveals that we are aware of properties of experiences as opposed to properties of the objects of experiences or other entities. This is a problem that has loomed large in philosophical discussions. Moore's observation concerning the

transparency of experience was advanced within a sense data account of experience. More recently, transparency has been used to support other views on the properties that typify conscious experiences. Different authors have employed the transparency consideration to support different views of the mutual relations of and, eventually, ontological priority of the phenomenal character of experiences and their representational content. (Kind 2010)

Instead, the more modest lesson that I derive from the transparency of experience is that introspective evidence supports the dependence of the self-ascription of a certain type of experience on the appreciation of certain facts. In having that type of experience, we can appreciate that something (i) *appears to be* a certain way and that (ii) there are certain *ways* in which this appears to be so. As an illustration for case (i), in having a color experience, we might be seeing that a certain object appears to be red. Cases of the type of (ii), involve ways in which we can see, for instance, that something is red. Consider the blurred way in which someone, who usually wears glasses, sees a red object when she is not wearing them.

It remains, thus, to be seen what is involved in the mastery of the concept of an experience of a certain type.

4. Compelling Transitions

First personal thoughts about our occurring conscious experience should rely on the capacities that allow us to think thoughts about how the world looks to us, including also the way in which we present the world as so appearing, when we have those experiences. Thus, we need to offer an account of the kind of transitions that lead us from experiencing the world to the judgement that we are having an experience of a certain type.

I maintain that the central transition the normative force of which any possessor of the concept of conscious experience of a certain type should be sensitive to, needs to have, for instance, the following form:

(1) I see in a certain way that x is F.

Then I can reach the judgement that:

(2) I am having the experience of the type involved in seeing in a certain way that x is F.

Such a transition should be supported by some principle that is relevant for possessing the concept of experience.

A proposal is that the principle would have to recommend a direct transition from seeing that something is the case, to the self-ascription of an experience of a certain type. This might suggest, from a psychological

point of view, that the capacities we might be employing in judging, let us say, that something is red are the same that enable us to think that we are having a certain experience of red.

This is not a plausible account. In general, from the judgment that something is red, it cannot follow that I am having an experience of red. Clearly, the transition seems more plausible when we judge that something is red based on what we are seeing. However, how the seeing that something is red can govern the transition to the thought that we are seeing something red? For sure, as I have argued in the previous section, we cannot become directly aware, in a quasi-perceptual fashion, of our color experience and thus determine that the experience has a certain typifying feature.

A plausible way out is represented by noticing that there are concepts, that we adopt in judging how things are in the world based on our experiences, that are different from concepts that might be employed in thoughts about the same facts that do not exploit the undergoing relevant visual experience. Let us consider, for instance, the difference between judging that a certain object has a square shape by seeing it and inferring the same judgment by being told that its shape has four 90 degree angles and equal sides. The application of the concept $SQUARE_1$, let us call it an *observational concept*, based on a visual discriminatory capacity appears to involve certain capacities, that are related to our seeing and having a certain type of experience, that are not involved in the concept $SQUARE_2$ that might be employed in the inferential reasoning.[7] That there is a difference between the observational concept $SQUARE_1$ and the non-observational one, might be proved with a hypothetical test of informativeness, that is based on the principle of conceptual distinctness that I have introduced above. An agent, would find informative that replacing in a propositional content the concept of a square whose possession conditions involve the ability to visually recognize squares, and the one that is spelled out by the description of a figure with four 90 degree angles.

The suggestion here is that the transition from judgments about how things are to self-ascriptions of the type of experiences that people have when they see things being that way, might rely, somehow, on the capacities that are involved in the mastery of the observational concepts. Thus, for example, a person who masters the concept of an experience of red should be able to rely on the visual capacities that enable her to judge that

---

[7] There is no tendency here to think that an ideal reasoner, for example with a complete understanding of geometry, could not infer a priori that something is $SQUARE_2$ from the fact that it is $SQUARE_1$. Thus, there is no tendency here to draw any analogy with the knowledge argument. Thanks to Marko Jurjako for raising this issue.

x is red, but also to judge that she is having a visual experience of a certain type.[8]

Of course, I am not arguing that the capacities involved in the application of the observational concept are sufficient for such a transition. However, they should surely be a prerequisite for the mastery of the concepts that enter thoughts the content of which is the self-ascription of a certain type of experience. Although, of course, many details need to be spelled out, it seems that assigning a role to the capacities to visually discriminate a certain feature in the mastery of the concept of conscious experience might bridge the gap in the inference from (1) to (2) above.[9]

Let me recapitulate what I have maintained so far. Investigating the content of thoughts that are involved in the self-ascription of types of conscious experience requires investigating the concepts that are implicated in these thoughts. The investigation of these concepts should rely on two general theoretical assumptions. First, two concepts are distinct if an agent might find informative their substitutions in a complete propositional content. Second, an account of a concept involves offering an account of its possession conditions. The knowledge argument by Frank Jackson can be interpreted as offering evidence for the conclusion that, for instance, the possession conditions of certain concepts of color experiences involve undergoing these conscious experiences. I have further argued that the explication of these possession conditions does not require the direct quasi-perceptual access to conscious experiences and their properties. Instead, I have suggested that the mastery of certain observational color concepts is a prerequisite for the use and possession of concepts about conscious color experiences.

Having clarified some features that any account of the mastery of phenomenal concepts should consider, we can now move on to the concept of the self that is involved in self-ascriptive thoughts of conscious experiences.

---

[8] I omit here the complication of handling self-ascriptive thoughts about having, for instance, a blurred visual color experience of an object that appears to be a certain color. But also in this case, I would invoke the central role of a recognitional concept of "blurriness." This concept, by relying on the ability to discriminate this feature when having that experience, would ground the specific way of thinking about the experience at issue.

[9] In Malatesti 2012, by elaborating Fred Dretske's displaced perception account of introspection (Dretske 1995), I explain the transition as based on connecting principles of the type "This object would not look blue to me unless I were having a color experience of a certain type." In that account, sensitivity to principles of that type is a necessary requirement for the possession of the concepts of conscious experiences.

5. In Favor of the Identification of the Self

The possession condition of phenomenal concepts explored so far discourages the view that in self-ascribing a certain type of conscious experience we exploit the direct awareness of the fact that our self has a certain conscious experience with a certain property. In fact, I have argued that we cannot be directly aware of color conscious experiences having certain discriminatory properties, and thus gain, directly, identifying information about them. However, it remains to be investigated the condition for the mastery of the concept "I," that is implicated in thoughts of the type "I am having a conscious experience of a certain type."

The investigation of the different senses that might enter in the specification of the concept expressed with the pronoun "I" is a complex issue.[10] Here I will just consider some aspects of that mastery that appears to be in harmony with the account of the possession conditions for the concept of type of conscious color experience that was sketched in the previous sections. In that account, it seems that the ground for the self-ascription of the relevant type of color experience should primarily be the exercise of the discriminatory ability in the application of the observational color concept. However, it seems that this proposal does not explicate completely the content of the self-ascriptive thought. How does the notion of the self as expressed by the pronoun "I" enter in the mastery of the concepts involved in the self-ascriptive thought?

It could be argued that the possession conditions of the concept "I" that enter in the mastery of the concept of conscious experiences, and thus the associate way of thinking of the self, could be cashed out in terms of information that is available just by having the specific experience. By visually discriminating the color of a surface, the subject might receive some self-specifying information that might enable the agent to have a certain way of thinking about herself. In fact, discriminating that a certain surface is or looks a certain color involves presenting the color as instantiated in a certain location in the egocentrically fixed space of the subject. Thus, also the information about the location of the self should be available when exercising the discriminatory capacity involved in the application of the observational concepts.[11]

---

[10] See Bermúdez 1998; Bermúdez forthcoming.

[11] This view is discussed under the heading of "situated subject account." (Gertler 2011: 226-232) This is an account that could be like the one advanced by Sidney Shoemaker in 1968. A noteworthy implication of accounts of this type is that our way of referring to the self does not involve identification information. This is also taken to render certain thoughts that involve that way of referring to the self immune to error through misidentification (IEM). Thus, at least in the case of the self-ascription of conscious

It seems however, that to master the notion of conscious experience and employ it in a self-ascription, we need a richer conception of the self than the one that might be merely available by having these experiences. We need a way to individuate and discriminate the self in a way that locates within it conscious experiences. Alan Millar has expressed this point well:

> Though we do not think of visual experiences as being in our eyes or auditory experiences as being in our ears it is plausible that it is constitutive of our notion of such experiences that they are typically obtained via, respectively, our eyes and our ears. Visual experiences are experiences of the sort which objects produce through their effects on our eyes and auditory experiences are experiences of the sort which sounds produce through their effects on our ears. (Millar 1996: 90)

Thus, the mastery of the concept of conscious experience involves the capacity to think about ourselves as entities that have sense organs and *internal* mental states that are determined by interactions with certain sorts of stimulation of these sense organs. Thus, the self-ascription of these mental states should involve, amongst other things, a way of thinking of our self as an entity of that type.

6. Conclusion

I have argued that self-ascriptive thoughts about color experiences, involve concepts that are individuated by means of sensitivity to two sorts of identification information. On the one side, there is identification information about the experience as the kind of state that enables us to see that something looks in a certain way. On the other side, that self-ascription involves thinking about the self by reference to specific sense organs that produce these internal states in virtue of certain stimulations. The characteristic ingredient of these self-ascribing thoughts about types of conscious experiences are concepts that involve modes of presentation of a specific way of looking of the world. These concepts are cognitively individuated by the discriminatory capacities that are conferred upon by having the types of experiences at issue.

---

experiences, if they manifest IEM, this cannot be due to the presence of the concept that is expressed with "I."

Acknowledgements

REFERENCES

Alter, T. and Walter, S. (2007). *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press.

Armstrong, D. M. (1968). *A Materialist Theory of the Mind*. London: Routledge & Kegan Paul.

Bermúdez, J. L. (1998). *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

Bermúdez, J. L. (forthcoming) *Understanding "I."* Oxford University Press.

Dretske, F. (1995). *Naturalizing the Mind*. MIT Press.

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

Gertler, B. (2011). *Self-knowledge*. Routledge.

Jackson, F. (1982). "Epiphenomenal Qualia." *Philosophical Quarterly 32*: 127–136. Reprinted in W. Lycan (ed.) *Mind and Cognition*. Blackwell 1990: 469-77.

Kind, A. (2010). "Transparency and Representationalist Theories of Consciousness." *Philosophy Compass* 902–913: 5-10.

Ludlow, P., Stoljar, D. and Nagasawa, Y. (2004). *There's Something About Mary*. MIT Press.

Malatesti, L. (2012). *The knowledge argument and phenomenal concepts*. Newcastle: Cambridge Scholars Publishing.

Millar, A. (1991). "Concepts, Experience and Inference." *Mind* 100, 4: 495–505.

Overgaard, S., Gilbert, P. and Burwood, S. (2013). *An introduction to Metaphilosophy*. Cambridge University Press.

Papineau, D. (2002). *Thinking about Consciousness*. Oxford: Clarendon Press.

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

Peacocke, C. (2009). "Concepts and possession conditions." In McLaughlin, B. P., Beckerman, A. and Walter, S. (eds.) *The Oxford Handbook of Philosophy of Mind*. Oxford University Press: 437-456.

Perry, J. (2001). *Knowledge, Possibility and Consciousness: The 1999 Jean Nicod Lectures.* MIT Press.

Putnam, H. (1967). "Psychological Predicates." In W. H. Capitan and D. D. Merrill (eds.) *Art, Mind and Religion*. University of Pittsburgh Press. Reprinted as "The Nature of Mental States." In H. Putnam *Mind, Language, and Reality. Philosophical Papers,* Vol. 2. Cambridge University Press (1975): 429-440.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson. Reprinted with an introduction by D. Dennett. London: Penguin (2000).

Shoemaker, S. (1968). "Self-reference and Self-awareness." *Journal of Philosophy* 65, no. 19: 555-567.

Shoemaker, S. (1996). *The First-Person Perspective and other Essays*. Cambridge University Press.

Smart, J. J. C. (1959). "Sensations and Brain Processes." Revised version printed in C. V. Borst (ed.) *The Mind/Brain Identity Theory*. London: Macmillan, (1970): 52-66.

Tye, M. (2000). *Consciousness, Color, and Content*. MIT Press.

Part III

SELF IN THE HISTORY OF PHILOSOPHY

6. The Logical Positivists on the Self
BORAN BERČIĆ

1. Introduction

Simon Blackburn starts his introduction to philosophy *Think* with a line: "We might say: it all began on 10 November 1619." (Blackburn 1999: 15) On that date Descartes allegedly had a vision and started writing his philosophical system. However, logical positivists did not share Blacburn's enthusiasm about Descartes' philosophy. Moritz Schlick was clear about it. Talking about *Cogito* as a candidate for the foundation of the whole human knowledge, Schlick said that: "Such a statement, which does not express anything itself, cannot in any sense serve as the basis of anything." (Schlick 1934: 218) He argued that it was a mere pseudostatement. Hans Reichenbach believed that *Cogito* "is one of the landmarks on the blind alley of traditional philosophy." (Reichenbach 1938: 261). No other philosophical movement ever criticized Descartes' *Cogito* so fiercely as logical positivists did. They criticized it on every occasion they could. (Schlick 1918: 85, 161; Carnap 1928: 261; Carnap 1932: 74; Schlick 1934: 218; Ayer 1936: 62, 187; Weinberg 1936: 184; Schlick 1936: 166; 184; Reichenbach 1938: 261; Von Mises 1939: 173; Reichenbach 1951: 35) This is understandable because they were radical empiricists. They firmly believed that no factual knowledge can be obtained *a priori*, by reason alone, and Descartes' *Cogito* was seen as a raw model of rationalistic philosophy, perhaps of philosophy in general. They all quoted 18th century German scientist and aphorist Georg Lichtenberg who said "*It thinks*, we should say, just as one says, *it lightens*. To say *cogito* is already too much, if we translate it as *I think*." (Lichtenberg 2012: 152; K 76) Although Lichtenberg was well known in the German speaking world, some authors believe that positivists quoted Lichtenberg because Ernst Mach did it in *The Analysis of Sensations*. (Mach 1886: 29; Blackmore 1972: 35; Williams 1978: 95) This is probably true because Mach really did have immense influence on the positivists. Since 1928. members of the Vienna Circle were institutionally organized in the *Verein Ernst Mach* (*Ernst Mach Society*).[1]

---

[1] I presented views of the logical positivists in *Filozofija Bečkog kruga (Philosophy of the Vienna Circle)* from 2002. This article is partly based on Chapter IX of the book.

2. Descartes' *Cogito*

There are four main ways to understand Descartes' *Cogito*.

(1) We can understand it as a sentence that expresses simple *awareness of our own existence*. This awareness is nonconceptal and noninferential. As soon as we think, we are aware that we think. And as soon as we are aware that we think, we are aware that we exist. According to this understanding, the awareness of our own existence is contained in the very act of thinking. One might say that this understanding is in the spirit of the movement of Phenomenology. However, this understanding of the *Cogito* is not very plausible. In Descartes' writings one cannot find sufficient support for it. It seems that this is not what Descartes had in mind.

(2) We can understand *Cogito* as a necessarily true *proposition* whose truth we grasp *a priori* by the insight of the reason. "One cannot think unless one exists." or "One who thinks has to exist." really seem like a good candidate for the *a priori* truth of reason. Also, there is a textual evidence for this interpretation. In *The Principles* Descartes talks about the eternal truths and says:

> We now come to speak of eternal truths. ... an eternal truth having its seat in our mind, and is called a common notion or axiom. Of this class are the following: It is impossible the same thing can at once be and not be; what is done cannot be undone; *he who thinks must exist while he thinks* [italics mine]; and innumerable others, the whole of which it is indeed difficult to enumerate, but this is not necessary, since, if blinded by no prejudices, we cannot fail to know them when the occasion of thinking them occurs. (Descartes 1644: XLIX)

However, in *Cogito* Descartes does not claim a general proposition that *whoever* thinks has to exist. He claims that *he* exists.

(3) Therefore, it is more plausible to understand *Cogito* as an *inference*. After all, it contains "therefore" and this indicates that it expresses an inference, not a single proposition. From the fact that he thinks Descartes *infers* that he exists. The proposition "he who thinks must exist while he thinks" should be taken as a hidden premise in the inference, not as a whole content of the *Cogito*. So, according to this interpretation, *Cogito* expresses the following inference:

P1: He who thinks must exist while he thinks.
P2: I think.
C: Therefore, I am.

This is certainly a sober interpretation that grasps well Descartes' intentions. However, it seems that the inference is more complicated and that it contains more hidden premises, in fact, a whole ontological theory. This

theory might be called *the S-A ontology*. The idea is that whatever exists is either a substance or an attribute. A substance is something that can exist on its own, something that does not need anything else for its existence. On the other hand, an attribute can exist only as an attribute of something distinct from itself, that is, as an attribute of a substance. Every substance has one essential attribute. The S-A ontology has a corresponding epistemology. Its central tenet is that we can be acquainted with a substance only through its attributes, we cannot be directly acquainted with a substance. And this is crucial in the discussion about *Cogito*. The picture is that once we are acquainted with the attributes, we *infer* the existence of an underlying substance to which these attributes belong. In *Principles* Descartes says:

> But yet substance cannot be first discovered merely from its being a thing which exists independently, for existence by itself is not observed by us. We easily, however, discover substance itself from any attribute of it, by this common notion, that of nothing there are no attributes, properties, or qualities: for, from perceiving that some attribute is present, we *infer* [italics mine] that some existing thing or substance to which it may be attributed is also of necessity present. (Descartes 1644: LII)

According to this interpretation, *Cogito* is an inference with several hidden premises of ontological nature: that thinking is an attribute and that an attribute has to belong to a substance. So, according to this picture, *Cogito* has to be reconstructed as follows:

1) There is thinking.
2) Whatever exits is either a substance or an attribute.
3) Thinking is an attribute.
4) Attribute must belong to a substance.
5) Therefore, there must be an Ego to which thinking belongs.[2]

The central characteristic of this picture is that *Ego* is not something that is directly observed but rather an *inferred entity*. Awareness of one's own existence is not an immediately given fact but rather *a product of theoretical reasoning*. In my opinion, this is the correct and full reconstruction of the *Cogito*. In the rest of the paper we will partly rely on this analysis.

(4) We can understand *Cogito* as a *performance*, that is, an utterance that is made true by the very act of uttering it.[3] *Cogito* is necessarily true in the sense that as soon as somebody says that he exists, it has to be true, it simply cannot be false. According to this understanding, *Cogito* is a nec-

---

[2] It is interesting to notice that this reconstruction of the argument does not start with "I think" but rather with the impersonal "There is thinking."
[3] Jaako Hintikka examines this interpretation in "*Cogito, Ergo Sum:* Inference or Performance" from 1962.

essary truth just like "I am here now." has to be true, no matter who, when and where says it. "I exist." is self-verifying, just as "I do not exist." is self-refuting. Nevertheless, although there certainly is an air of performativity in the *Cogito*, we cannot say that this is what Descartes had in mind. His writings do not support this interpretation. In the exposition of *Cogito*, Descartes puts stress on other things, not on its self-verifying character.

3. The Logical Positivists on the *Cogito*

3.1. Moritz Schlick: *Cogito* is a Stipulation

In the *General Theory of Knowledge* from 1918, in §12 *What Knowledge is Not*, Moritz Schlick argued that *Cogito* is not a statement (that can be true or false) but rather a stipulation, or a concealed definition:

> Certainly the judgment *"cogito, ergo sum"* (after all necessary corrections are made) does express an incontrovertible truth, namely, that content of consciousness exist. But we saw some time back that not every truth need be knowledge; truth is the broader concept, knowledge the narrower one. Truth is uniqueness of designation, and uniqueness can be obtained not only through knowledge, but also through definition. And this is the case here. Descartes' thesis is a *concealed definition* [italics mine]; it is an improper definition of the concept *existence* - what is earlier called a "concrete definition." What we have is simply a stipulation that experience, or the being of contents of consciousness, is to be designed by the words "ego sum" or "the contents of consciousness exist." (Schlick 1918: 85)

To understand *Cogito* as a stipulative definition might seem like an interesting idea but obviously it cannot serve Descartes' purposes. To serve the purpose of the Archimedean point of knowledge, *Cogito* cannot be a stipulative definition true by *fiat*, it has to be understood as a statement that expresses its objective truth makers. It is doubtful that in *Cogito* Descartes introduces and defines the concept of existence. It rather seems that he has previous and independent understanding of that concept and that he applies it in the *Cogito*. Generally, it is a very interesting question of how much one has to know to come to the *Cogito*. Obviously, one has to have a mastery of some concepts and principles of thought. It would be unfair to argue that Descartes introduced the concept of existence in *Cogito*. In *Principles*, paragraph X, he says what is needed to arrive to *Cogito*:

> When I have said that this proposition, I THINK, THEREFORE I AM, is the first and most certain one encountered by anyone who conducts his thinking in an orderly manner, I have not, however, said that it was not necessary to know aforehand what thinking, certainty and existence are, and that in order to think one must be, and other such similar matters; but because these notions are so simple that, by themselves, they do not make us aware of anything that exists, I have not deemed it necessary to give an account of them here. (Descartes 1644: X)

Tabula rasa cannot arrive at *Cogito*. Remember, in Descartes' epistemology, the belief that I exist is not the first belief that we have, it is the first *justified* belief that we have. Nevertheless, Schlick has more to say about *Cogito*. In *General Theory of Knowledge*, §20, named *So-Called Inner Perception*, he says:

> The Cogito of Descartes, as we remarked earlier, contains the trap of a distinction between a substantivist "I" and its activity, into which Descartes fell when he added: *ergo sum*. For as is easily seen, his *sum* means for him the existence of a substantial "I." Lichtenberg's very true observation that Descartes should have said "It thinks" instead of "I think", is not only an inspired remark but should really be made the supreme guiding principle of psychology. ... The stream of consciousness is simply an existing process; the "I" is the unified interconnection of this process, not a person who inspects and guides it. (Schlick 1918: 161)

As we saw earlier, Descartes believed that the inference from "there is thinking" to "there is somebody who thinks" is assured by the *common notion* or *axiom* of the S-A ontology. The relevant common notion is that "from perceiving that some attribute is present, we infer that some existing thing or substance to which it may be attributed is also of necessity present." (*Principles*, LII) On the other hand, as we can see from this quotation, Schlick, together with Lichtenberg and others, believed that this inference is nothing but a *logical fallacy* of substantivisation (or hypostatization, or reification). Now, what we have here, an axiom of reason or a logical fallacy?

Schlick's argument can be seen as an instance of a wider philosophical discussion: the empiricist critique of the rationalist conception of substance. Empiricists are proponents of the *bundle theory of substance*, where a substance is seen simply as a bundle of properties without any underlying substratum to which these properties are supposed to be attached. Rationalists, on the other hand, accept the *substratum theory of substance* and argue that every substance is composed of properties and a substratum to which these properties belong. For empiricists a thing is nothing but a bundle of properties, while for rationalists a thing is a bundle of properties attached to their carrier, that is, to a substratum. In the case of the *Cogito* argument, Cartesian *Ego* is the substratum. Schlick, as empiricists, rejects the idea of an underlying occult entity and, as we saw, argues that "I" is nothing but "the unified interconnection ... of the stream of the consciousness." (Schlick 1918: 161) There is no underlying entity to which this stream belongs, there is no homunculus "who inspects and guides it." Roughly speaking, Schlick defends *a bundle theory of the self.*[4] Though, we

---

[4] I say "roughly" because Schlich argues that Hume's bundle theory cannot account for

have to note here that Cartesian S-A ontology with its substratum theory of substance is not the only ontological framework in which we can infer "there is somebody who thinks" from "there is thinking." After all, Schlick in the very same paragraph explains what "I" stands for. Within the framework of the bundle theory of substance one can also infer that "there is somebody who thinks" from "there is thinking." The only thing that is needed is a plausible assumption that properties always come in bundles. In the case of *Cogito* this assumption amounts to the claim that psychological processes always take place in the corresponding bundles, that is, in the human selves. It seems that Lichtenberg simply went too far here. To eliminate occult Cartesian *Ego* from the ontology is one thing, but to claim that thinking can occur without a person who thinks is another thing. The first claim is plausible, the second one is not. Cartesian inferences might be valid without his ontology. We can say "I think" and "I am" without commitment to substratum theory of substance and its occult entities.

3.2. Rudolf Carnap: *Cogito* is Meaningless because it cannot be Formulated in the Language of Logic

In "The Elimination of Metaphysics through Logical Analysis of Language," a programmatic article from 1932, Rudolf Carnap eliminates Descartes' *Cogito* as a metaphysical piece of nonsense, on pair with Heidegger's "Nothing nothings." or Hegel's "Pure Being and pure Nothing are, therefore, one and the same." Carnap's objections to *Cogito* here are not substantial, but rather formal. In his opinion, *Cogito* is ungrammatical and it cannot even be formulated in a decent language. Although the grammar of natural languages allow formulation of such a sentence, the logical grammar forbids it. Talking about *Cogito*, Carnap says:

> We notice at once two essential logical mistakes. The first lies in the conclusion "I am." The verb "to be" is undoubtedly meant in the sense of existence here; for a copula cannot be used without predicate; indeed, Descartes' "I am" has always been interpreted in this sense. But in that case this sentence violates the above-mentioned logical rule that existence can be predicated only in conjunction with a predicate, not in conjunction with a name (subject, proper name). An existential statement does not have the form "a exists" (as in "I am", i.e. "I exist"), but "there exists something of such and such a kind." The second error lies in the transition from "I think" to "I exist." If from the statement "P(*a*)" ("*a* has the property P") an existential statement is to be deduced, then the latter can assert existence only with respect to the predicate P, not with respect to the subject *a* of the premise. What follows

---

the *unity of consciousness*. (Schlick 1918: 123) Schlick dedicates a whole paragraph to that problem - §17 *The Unity of Consciousness*. I will not go deeper into this problem here.

from "I am European" is not "I exist", but a "a European exits." What follows from "I think" is not "I am" but "there exists something that thinks. (Carnap 1932: 74)

The first mistake that Carnap talks about is that the verb "to be" is used in two senses, as a copula and as a predicate. However, the argument runs, existence cannot be used as a predicate. In fact, this is old Kant's critique of the ontological argument for the existence of God: existence cannot be a predicate. And Carnap mentiones that on the same page. Though, we do talk about particular things that do or do not exist. We say that Kraljević Marko really existed or that Atlantis never existed.[5] And we do not have any problems understanding the meaning of these claims. The second mistake that Carnap talks about is that "I am" does not follow from "I think." What follows from "I think." is "There exists something that thinks." As we saw, Descartes justified the inference from "I think." to "I exist." with the eternal truth (or common notion or axiom) that *he who thinks must exist while he thinks*. Would this be sufficient to infer "I am." from "I think."?

Here we have to have in mind that logical positivists took logic very seriously. At many places they argued that natural language is faulty in many ways, that it is imprecise and misleading. For them the idea of a perfect language seemed natural and fruitful. They believed that traditional philosophical problems are nothing but logical mistakes, and that careful logical analysis would solve them all. Moreover, they believed that traditional philosophical problems are pseudoproblems that cannot even be formulated within the framework of the ideal language of the contemporary logic. After all, Carnap believed that philosophy is nothing but a *logical syntax of the language of science*. For these reasons, logical positivists took very seriously this objection to *Cogito*. Now, assuming that *Cogito* really cannot be formulated in the language of the first order predicate logic, in principle we can react in two opposite ways. We can argue, as Carnap did, that the language of contemporary logic is the best language we have and that we should reject as illegitimate anything that cannot be formulated in it. Or, we can argue that *Cogito* is perfectly legitimate and meaningful, and that therefore there must be something wrong with the contemporary logic if something so simple and understandable like *Cogito* cannot be formulated in it. If contemporary logic cannot accommodate *Cogito,* so much worse for the contemporary logic. Here we can quote Wittgenstein's comment from *Philosophical Investigations* on the relationship between the ideal and the actual language:[6]

---

[5] Kraljeveć Marko is a heroic character from the medieval oral literature.
[6] Majda Trobok pointed this out to me.

> We have got on to slippery ice where there is no friction and so in a certain sense the conditions are ideal, but also, just because of that, we are unable to walk. We want to walk: so we need friction. Back to the rough ground! (Wittgenstein 1953: §107)

3.3. Julius Rudolph Weinberg: *Cogito* is a Valid but Empty Inference

Julius Weinberg in his book *An Examination Of Logical Positivism* from 1936 accepts Carnap's argument and makes an interesting comment about it. Weinberg argues that *Cogito* can be interpreted as a valid inference, but under that interpretation it would be a tautology, deprived of any factual content and as such it could not serve Descartes' purposes.

> "Something thinks" implies "something thinking exists." This, in logical symbolism, is $\varphi u \ . \ \supset \ . \ (\exists x)\varphi x$, which is a tautology. Tautologies assert no facts because, as has been shown above (Chapter II), they are entirely concerned with symbols. In this case φu is one way of saying $(\exists x)\varphi x$. Nothing has been demonstrated about the world. On this hypothesis, the cogito is a deduction but it presents nothing new, and, moreover, does not demonstrate what Descartes attempted, i.e. that a simple, identical, substantial, and spiritual entity exists. The important thing to notice about this treatment of the cogito is the elimination of the first person from the proposition. The means of determining the sense of "I think" cannot be given, so that, in this form, the proposition is meaningless, whereas if it is changed to "something thinks", the deduction "a thinking thing exists" is evidently no new information. Consequently nothing metaphysical could be intuited or inferred from the proposition. (Weinberg 1936: 184)

Perhaps the most interesting part of Weinberg's comment is the claim that *Cogito*, if understood in the sense of "I think." is meaningless because "The means of determining the sense of 'I think.' cannot be given." Maybe this was Carnap's real motive, but, as we saw, this was not his claim. His claim was that "I am." does not follow from "I think.", not that we cannot determine the sense of "I think." Maybe I am going too far here but it seems that Weinberg's worry was partly extra-logical. The claim that *Cogito* cannot be formulated in the language of the contemporary logic is one thing, while the claim that we do not really understand what it means is another.

3.4 Alfred Jules Ayer: "I exist" does not follow from "There is a thought now"

Language, Truth and Logic from 1936 is regarded as a book that brought logical positivism into the Anglo-Saxon world. Ayer opens Chapter 2 *THE FUNCTION OF PHILOSOPHY* with the claim that one of the superstitions about philosophy is that "the business of philosophy is to construct a deductive system." (Ayer 1936: 62) The paradigmatic case of such a system is Descartes' philosophy. Here is what Ayer says about it:

> What he was really trying to do was to base all our knowledge on proposi-
> tions which it would be self-contradictory to deny. He thought he had found
> such a proposition in "*cogito*", which must not here be understood in its
> ordinary sense of "I think", but rather as meaning "there is a thought now."
> In fact he was wrong, because "*non cogito*" would be self-contradictory only
> if it negated itself: and this no significant proposition can do. But even if it
> were true that such a proposition as "there is a thought now" was logically
> certain, it still would not serve Descartes's purpose. For if "cogito" is taken
> in this sense, his initial principle, "*cogito ergo sum*", is false. "I exist" does not
> follow from "there is a thought now." The fact that a thought occurs at a giv-
> en moment does not entail that any other thought has occurred at any other
> moment, still less that there has occurred a series of thoughts sufficient to
> constitute a single self. (Ayer 1936: 62, 63)

Ayer has two arguments here. The first one is that *Cogito*, understood as
"There is a thought now." is not necessary. The second one is that "I exist."
does not follow from "There is a thought now." Let's focus on the first ar-
gument. Of course, it is questionable whether the first part of the *Cogito*
should and could be understood as "There is a thought now." instead of "I
think." Though, we have to say that Ayer is benevolent here, he looks for
the formulation that might serve Descartes' purposes, that is, the formu-
lation that would be impossible to deny. Ayer's point is that, contrary to
Descartes' views, "There is a thought now." can be denied without contra-
diction. "There is no thought now." is not a contradiction, just like "There is
a thought now." is not a tautology. It is simply a contingent matter whether
there exists a thought now or not. On the one hand, this analysis is correct,
it really is a contingent matter whether there are any thoughts at this mo-
ment. A universe without thoughts is not a contradiction. It is a consistent
idea. But one the other hand, as we saw at the beginning of this article,
there is an air of self-verifying performance in the *Cogito*. If at this moment
somebody would think a thought "There is a thought now." his thought
would be necessarily true. The very act of thinking it would make it true.
The situation is analogous to the following one. If the sentence "Something
is written on this wall!" is written on this wall, then it is self-verifying and
necessarily true. If it is uttered by someone who points to the wall, then it
is contingently true or false, depending on whether something is written
on the wall or not.[7] So, although performative character of the *Cogito* was
not in the focus of the Descartes' argumentation, there is a sense in which
"There is a thought now!" is necessarily true. The second Ayer's argument is
that "I exist." does not follow from "There is a thought now." Ayer believes
that we are dealing with a *non sequitur* here because "a series of thoughts"
is needed to constitute a self and we have only a single thought. For Des-

---

[7]   This is the difference between the semantic and the pragmatic paradox.

cartes a single thought is sufficient to get the *Cogito* off the ground. A single thought, in conjunction with the axiom that "He who thinks must exist while he thinks," entails that there is somebody who thinks. Also, under the assumption of S-A ontology, the occurence of a single thought entails that there is somebody who thinks. If I add the premise that I can think only my own thoughts (not thoughts of other people), I have a right to infer that I think. Of course, Ayer does not rely on the Descartes' axiom, nor on the S-A ontology. He accepts a kind of the bundle theory of the self and for him a single thought is not sufficient to infer that he exists. He needs a whole series.

4) What is the Self?

In the previous chapter we presented a critique of *Cogito*. That was a negative part of the positivists' views about the self. However, they had a very interesting and quite elaborated positive part as well. They tried to say what self is.

4.1. Rudolf Carnap: The Self is the Class of Elementary Experiences

In the *Aufbau* Carnap defined self in §163 *The Problems of the Self*:[8]

> *The "self" is the class of elementary experiences.* It is frequently and justly emphasized that the self is not a bundle of representations, or experiences but a unit. This is not in opposition to our thesis, for (as we have shown in §37 and have emphasized repeatedly) a class is not a collection, or the sum, or a bundle of its elements, but a unified expression for that which the elements have in common. (Carnap 1928: 260)

Carnap was well aware of the old objection to the bundle theory of the self. It is not sufficient to say that we are a bundle of experiences. A satisfactory analysis of the self has to grasp the fact that our experiences have a kind of unity. Carnap argued that the concept of a class is the right concept for this task because a class is a "unified expression for that which the elements have in common." But it is questionable whether a concept of a class can really provide a kind of unity that is needed here. Take for instance a class of people taller than 1.80m. The only thing that they have in common is the fact that they are taller than 1.80m. They do not have a kind of unity we believe our experiences have. In the same way, the only thing that elements of the class of elementary experiences have in common is the fact that they are elementary experiences. And this fact alone certainly cannot provide the kind of unity that we are looking for here. The fact that they are elementary experiences cannot tell us that they stand in different relations;

---

[8] In the German original Carnap talks about *das "Ich."*

that they have spatial and temporal order, causal order, that they can be used in explanations or inferences, that they have characteristics of a functional unity, etc.[9] So, it seems that the concept of a class, by itself, cannot provide a unity of consciousness. Nevertheless, let's take a further look at the Carnap's proposal. Carnap defines class in §33. *Classes*. He says:

> The extension of a propositional function with only one argument position, i.e., the extension of a property, is called a *class*. ... Classes, since they are extensions, are quasi objects. Thus the class symbols do not have independent meaning; they are merely aids for making statements about all the objects which satisfy a given propositional function without having to enumerate them one by one. Thus the class symbol represents, as it were, that which these objects, i.e., the elements of the class, have in common. (Carnap 1928: 57)

Philosophy is supposed to unveil deep and important truths about ourselves. We expect philosophers to tell us what we really are, or what is our deepest nature, what is the meaning of life, etc. At least we expect philosophers to tell us something about the *condition humaine*. Having this in mind, Carnap's definition might sound like a joke. He tells us that we are "extensions of propositional functions." We are neither rational animals, nor featherless bipeds, nor thinking things. We are extensions of propositional functions! And this is what we really are! This is our ultimate nature! But what sense does it make? How can we be logical entities? Well, this does not mean that we are logical entities. To say that an object can be described mathematically is one thing, and to say that an object is a mathematical object is quite another thing. Trajectories of celestial bodies can be described mathematically, but this does not mean that celestial bodies *are* mathematical entities. They are mostly rocks. Now, since we are classes of elementary experiences, and elementary experiences are psychological entities, one might conclude that we as well are psychological entities. However, things are not so simple. Classes need not and can not have properties that their elements have. The class of wooden objects is itself not a wooden object, the class of rectangular objects is itself not rectangular, etc. Carnap is explicit about it:

> Not only is it not the case that a class is identical with the whole corresponding to it; it even belongs to a different sphere. ... *Nothing can be asserted of a class that can be asserted of its elements.* ... a class does not belong to the same sphere as its elements. (Carnap 1928: 64)

So, although experience is the stuff that we are made of, we are not experience, we belong to a different domain. Now, the question that we might ask here is whether Carnap was a reductionist or antireductionist about the

---

[9] Not to mention the stronger claim that they are *ours*, that is, that they belong to a single conscious subject.

self. What was his view, that I am *nothing but* my experience, or that I am *something over and above* my experience? On the one hand, he obviously was a reductionist about the self. *Aufbau* was essentially a reductionistic project. In a preface to the second edition he says that the central thesis of the book is that "it is in principle possible to reduce all concepts to the immediately given." (Carnap 1928: vi) Since everything else is reducible to the immediately given, so is the self. Also, in a §33 quoted above he says that "the class symbols do not have independent meaning; they are merely aids for making statements about all the objects which satisfy a given propositional function without having to enumerate them one by one." (Carnap 1928: 57) Let me paraphrase this statement. It means that the pronoun "I" does not have independent meaning but that it is merely an aid for making statements about all the elementary experiences I have without having to enumerate them one by one.[10] In other words, when I talk about myself, I in fact talk about all of my elementary experiences. There is no special entity that I talk about. Carnap claims that a class symbol "by itself means nothing." Talking about the class symbol *"ma"* (of a propositional function "*x* is a man.") he says: "Even though *ma* itself does not designate anything, one speaks of "the designatum of *ma* as if it were an object." (Carnap 1928: 58) This is a very strong reductionistic claim. However, on the other hand, Carnap also makes antireductionist claims about the self. As we saw above, he argues that we cannot assert of the class the same things that we can assert about its elements, and that classes and their elements belong to a different spheres. In §37 *A Class Does Not Consist of its Elements* Carnap says: "*Classes cannot consist of their elements as a whole consists of its parts. Classes are quasi objects relative to their elements; they are autonomous complexes of their elements.*" (Carnap 1928: 63) So, to paraphrase, we are quasi objects relative to our elementary experiences, or, we are autonomous complexes of our elementary experiences. And this is a very strong antireductionist claim. Also, Carnap quotes Frege who said "The extension of a concept does not consist of the objects which fall under the concept." (Carnap 1928: 64)

Now, the question is whether Carnap is a reductionist here or an antireductionist. Obviously, he has inclinations for both options. But the question is whether his views are consistent. Can he have a pie and eat it? The

---

[10] Carnap's view has one flaw. Since we are *classes* (the extensions of a propositional function with only one argument position), the basis of reduction is necessarily limited to only one kind of things (elementary experiences). This means that *body* cannot be included in the basis of reduction. As we will see, in this respect Reichenbach's *abstracta* are much more plausible candidates because they can be composed of different kinds of things.

general problem with the reductionism and antireductionism about the self is that, on the one hand, it seems that reductionism is not enough, while, on the other hand, it seems that antireductionism is too much. On the one hand, we are inclined to think that we are something that *has* experience (not that we just *are* experience). On the other hand, we do not want to postulate the existence of Cartesian Egos, bare particulars, substrata, or other occult entities. And it seems that this is exactly what the concept of a class provides. On the one hand, a class is not reducible to its elements, while, on the other hand, there is no special entity to which it refers. We might say that the introduction of the concept of a class enabled Carnap to defend *conceptual antireductionism* and *ontological reductionism*. In other words, it enabled him to navigate between the Scylla of reductionism and the Charybdis of antireductionism. If we have to make an overall verdict on whether Carnap was a reductionist or an antireductionist about the self, I think that we should say that, all things considered, he was an antireductionist about the self. The main reason for this verdict would be the fact that on many places in the *Aufbau* he insists on the point that classes are not reducible to their elements.

In trying to decide whether Carnap was a reductionist or an antireductionist about the self, perhaps one more thing might be relevant. It is a general question whether there is any reality behind the objects that he talks about. However, he systematically refuses to answer this question. He rejects it as meaningless. In §5 *Concept and Object*, he says:

> Does thinking "create" the objects, as the Neo-Kantian Marburg school teaches, or does thinking "merely apprehend" them, as realism asserts? Construction theory employs a neutral language and maintains that objects are neither "created" nor "apprehended" but *constructed*. I wish to emphasize from the beginning that the phrase "to construct" is always meant in a completely neutral sense. From the point of view of constructional theory, the controversy between "creation" and "apprehension" is an idle linguistic dispute. (Carnap 1928: 10)

Here we should rely on the distinction that Carnap explicitly introduced later. (Carnap 1950) It is the distinction between *internal* and *external* questions. If the question whether Carnap believed that selves really exist is understood as a question internal to the constructional system of the *Aufbau*, the answer is positive. Yes, he believed that selves exist! They are constructed and they exist! However, if the question is understood as external to the system, then the answer is that he rejected the question as meaningless.

4.2. Alfred Jules Ayer: The Self is a Logical Construction out of Sense-Experiences

In *Language, Truth and Logic* from 1936, *Chapter 7 THE SELF AND THE COMMON WORLD*, A. J. Ayer says what the self is:

> We know that a self, if it is not to be treated as a metaphysical entity, must be held to be a logical construction out of sense-experiences. It is, in fact, a logical construction out of the sense-experiences which constitute the actual and possible sense-history of a self. And, accordingly, if we ask what is the nature of the self, we are asking what is the relationship that must obtain between sense-experiences for them to belong to the sense-history of the same self. And the answer to this question is that for any two sense-experiences to belong to the sense-history of the same self it is necessary and sufficient that they should contain organic sense-contents which are elements of the same body. (Ayer 1936: 165)

This analysis is in the spirit of Hume's *bundle theory of the self*. However, Ayer warns us that there is an important difference. In Hume's analysis self is a *bundle* or *aggregate* of experiences, while in Ayer's analysis self is a *logical construction* out of experiences. Now, the question here is what is a logical construction.[11] *X* is a logical construct out of *a*, *b* and *c* if and only if sentences about *X* can be translated into sentences about *a*, *b* and *c*. "What we hold is that the self is reducible to sense-experiences, in the sense that to say anything about the self is always to say something about the sense-experiences." (Ayer 1936: 168) Of course, it is questionable whether such program can really be carried out.[12] Hume had a problem; he did not know how to prove that two experiences belong to the same self. Ayer offers a solution here. Roughly speaking, the answer is that they belong to the same *body*. Ayer also offers a solution to the problem of *epistemic subject*. A problem for any version of the bundle theory is that experiences have to belong to a subject, they cannot be subjectless. Experience has to be *somebody's* experience! And this is the problem for the reductionism about the self. How can the self be constructed out of experience when the very notion of experience presupposes a self to which it belongs? Ayer agrees

---

[11] Logical positivists took this notion from Russell and used it extensively. Carnap starts his *Aufbau* by quoting Russell. "The supreme maxim in scientific philosophizing is this: Whenever possible, logical constructions are to be substituted for inferred entities." (Carnap 1928: 5; Russell 1914: 155)

[12] In fact, this ambition amounts to replacing *personal* language with the *impersonal* one. But the question is whether a complete impersonal description of the world would be a complete description of the world. The worry is that it would miss something very important; that I am BB, that you are ... , etc. The issue was raised by Nagel (1986). Very nice exposition, as well as contribution, to the debate can be found in Baker (2013). Although this is a very important issue, I will not discuss it here.

that experience has to belong to a subject, but he does not believe that this forces us to stipulate the existence of the Cartesian mental substance. He tells us how we can think and talk about the epistemic subject without commitment to a suspicious metaphysical baggage.

> We shall see that this relation of being experienced by a particular subject is to be analysed in terms of the relationship of sense-contents to one another, and not in terms of a substantival ego and its mysterious acts. (Ayer 1936: 161, 162)

This approach to the analysis of a subject is not only ontologically more economic. We can pay the ontological price if we have to. The point is that this kind of analysis is methodologically far superior to the Cartesian analysis. To say that we can think because we are thinking things is to explain nothing. It is a raw model of *virtus dormitiva* explanation. The reductive analysis of the self is intrinsically more fertile because it explains characteristics of the self as relationships between the elements, not as its primitive characteristics. If we introduce, say, second order desires or higher order thoughts, we can explain something about ourselves. But what could we explain if we introduce a substance whose essential attribute is thinking?

In his critique of *Cogito* Ayer does not rely on the logical analysis only. He also relies on the assumptions of empiricism, verificationism and neutral monism. As empiricists, Ayer hailed Locke's famous critique of the notion of substance as something "we know not what" that supports and holds together observable properties of material objects. (Locke 1690: 269; Book II, Chapter XXIII, §3) Though, Ayer believed that the same holds for the Cartesian notion of mental substance. No matter whether substance is physical or mental, we have no reason to stipulate its existence.

> For it is clearly no more significant to assert that an "unobservable somewhat" underlines the sensations which are the sole empirical manifestations of the self than it is to assert that an "unobservable somewhat" underlines the sensations which are the sole empirical manifestations of a material thing. (Ayer 1936: 166, 167)

Generally speaking, logical positivists did not rely on the principle of verification in their rejection of *Cogito*, as one might expect. They primarily treated *Cogito* as a logical error and dismissed it on *a priori* grounds. Ayer is also explicit about it. Nevertheless, in a couple of places he criticizes Cartesian argumentation from a verificationist perspective. The assumption that there is a mental substance is not "capable of being verified." (Ayer 1936: 161) Also, immortal soul is a "metaphysical entity, concerning which no genuine hypothesis can be formulated." (Ayer 1936: 168)

One of the shared assumptions of logical positivism was *neutral monism*.[13] It is the view that basic constituents of knowledge are neither physical nor mental, but rather neutral with respect to this distinction. The idea is that physical and mental has to be constructed out of these neutral elements. Basic elements are, by themselves, not yet physical or mental.

> And we have seen that the terms "mental" and "physical" apply only to logical constructions, and not to the immediate date of sense themselves. Sense contents themselves cannot significantly be said either to be or not to be mental. (Ayer 1936: 187)

For Ayer, these basic elements are *sense-contents*. Obviously, neutral monism provides a very good platform for the critique of *Cogito*. Since my own mind is also a construct out of the basic and neutral elements, I cannot be sure about the content of my own mind and doubt everything else.[14] Just as it is logically possible that physical objects do not exist, it is logically possible that mental objects do not exist. Ayer says that Berkeley was right when he offered a phenomenalistic analysis of physical objects, but wrong when he did not offer such analysis of mental objects. (Ayer 1936: 167) For this reason, idealism, solipsism and *Cogito* are ill formed. And it was Descartes who was also guilty of this error, so influential in the history of western thought. In the concluding chapter of *Language, Truth and Logic*, *Chapter 8 SOLUTIONS OF OUTSTANDING PHILOSOPHICAL DISPUTE*S, Ayer says:

> I think that the idealist view that what is immediately given in sense-experience must necessarily be mental derives historically from an error of Descartes. For he, believing that he could deduce his own existence from the existence of a mental entity, a thought, without assuming the existence of any physical reality, concluded that his mind was a substance which was wholly independent of anything physical; so that it could directly experience only what belonged to itself. (Ayer 1936: 187)

Things are clear here. If neutral monism is right, Descartes has to be wrong. If the basic elements of our knowledge are neutral, then it cannot be true that *Cogito* is "the first and most certain thing to occur to anyone who philosophizes in an orderly way." (Descartes 1644: 2, 3; §10) A chain of epistemic justification cannot start with *Cogito*. Before that we have to construct *I* and *thinking* out of neutral elements. However, even if we accept this analysis, the interesting question is whether we can proceed with Cartesian epistemology and doubt the existence of the world *once* we construct *I* and *thinking* out of neutral elements. Can we consistently assert the following two propositions?

---

[13] Logical positivists inherited this view primarily from Mach and Russell.
[14] Perhaps Lichtenberg dictum should also be understood in this sense.

(1) *I* and *world* are constructed out of neutral elements.

(2) I can doubt the existence of the whole world but I cannot doubt my own existence.

Perhaps (1) and (2) are not in a direct contradiction, but there certainly is some tension between them. Here we have another pair of propositions:

(1) I can develop a concept of a self only if I have a body.

(2) I can doubt whether I have a body.

The idea is that *once* I develop a full concept of a self, I can consistently doubt whether I really have a body. Of course, the question is whether this is consistent.

4.3. Hans Reichenbach: The Ego is an Abstractum Composed of Concreta and Illata

In *Experience and Prediction* from 1938, in §28. *What is the Ego?* Hans Reichenbach says what the ego is, that is, what is the thing that "I" refers to:

> The ego is an abstractum, composed of concreta and illata, constructed to express a specific set of empirical phenomena. ... First is the fact that among all human bodies there is one, our own body, which accompanies all phenomena. ... There is, second, the fact that some physical phenomena are known to ourselves alone. ... We find in this way that our description of the physical world differs in some respect from the description of other people. The set of facts we refer to here is the same as expressed by the idea that the immediate world is directly accessible to one person alone. It is the whole of these facts which is comprehended by the abstractum "ego." (Reichenbach 1938: 259, 260)

Here we have to explain what abstractum is. For Reichenbach, "abstract" does not mean "out of space and time," as it is often used today. In his ontology Reichenbach has three kinds of entities; *abstracta, concreta* and *illata*. *Concreta* are middle sized physical objects that we encounter in the world; chairs, tables, cats, etc. *Illata* are inferred entities; atoms, mental states, etc. *Abstracta* are entities that are constructed out of concreta; "political state, the spirit of the nation, the soul, the character of a person." (Reichenbach 1938: 93; §11. *The existence of abstracta*) Now, the question is whether abstracta exist, more precisely, in this context the relevant question is whether abstracta exist *on their own*, or they are reducible to concreta without remainder. In Reichenbach's opinion, abstracta do not have *per se* existence, they are completely reducible to concreta. "To one abstract proposition we co-ordinate a group of concrete propositions in such a way that the meaning of the group is the same as the meaning of the abstract proposition." (Reichenbach 1938: 95) Since an abstract fact can be realized in more than one way, a reductive proposition will be a disjunction of conjunctions. (Re-

ichenbach 1938: 95) Derek Parfit would say that for Reichenbach there was "no further fact" about our own existence. But still, the question is whether Reichenbach was a reductionist or an eliminativist about the self. If the self is reducible without a remainder, what does it mean? That it exists or that it does not exist? In a manner of a good logical positivist, Reichenbach argues that it is a pseudoquestion:

> We see, then, that the question whether or not abstracta exist, whether or not there is the term only or also a corresponding entity, is a pseudo-problem. The question is not a matter of truth-character but involves a decision - a decision concerning the use of the word "exist" in combination with terms of a higher logical order. ... The decision may even depend on the profession of the speaker. For a merchant supply and demand may be existent entities, whereas an electrician would conceive an electrical charge as existent. It is a remarkable psychological fact that this "feeling of existence" which accompanies certain terms is fluctuating and depends on the influence of the milieu. The pursuit of this question is of great psychological interest; for logic there is no problem at all. (Reichenbach 1938: 97)

But, are we abstracta? Do we really belong to the same ontological category as supply and demand, race, or spirit of the nation? In *Categories* Aristotle claimed that we are *primary substances*, a paradigmatic case of existence. (Aristotle 1963: 5; 2a11) In *Individuals* P. F. Strawson argued that we are *basic particulars*. (Strawson 1959: 38) Although we know that we have parts, in more than one sense, we mostly think and talk about ourselves as *individual substances*. Reichenbach's claim that we are abstracta seems just false. Where did he go wrong? If he did, of course. It seems that Reichenbach's analysis of reduction of a complex to its internal elements does not take into account *the level of integration* of complexes. Different complexes have different levels of integration. It makes some sense to claim that Indonesia does not really exist and that what really exists are 18 thousands islands. It makes much less sense to claim that Australia does not exist and that what really exists is its eastern half and its western half. The difference is in the obvious fact that Australia is territorially much more integrated than Indonesia. Different parts of a single man stick together much more firmly than different parts of a nation or of a race. And this is why a man is a much better candidate for a really existing entity than a nation or a race. Although a general reductionistic schema "*X* is nothing but *a, b, c, ...*" or "*X* is nothing over and above *a, b, c, ...*" can be satisfied by different candidates, it does not mean that we should categorize all of these candidates as abstracta.[15]

---

[15] Perhaps Reichenbach's distinction between abstracta and concreta should be understood as *relational*; that elements of an abstractum are concreta *in relation* to that abstractum.

This is Reichenbach's "official view" about the ontological status of the self in *Experience and Prediction*. Though, perhaps there is something puzzling in his writings. Sometimes he talks about the *construction* of the Ego, sometimes about the *discovery* of the Ego. However, the expression "discovery of the *X*" implies realistic construal of the *X*. It implies that *X* is something that exists before and independently of our discovery of it. Of course, within the framework of the positivistic *constructional system* the difference between the construction and the discovery fades away. After all, physical objects (concreta) are also constructs, they are constructed out of impressions. Nevertheless, when one goes through the Reichenbach's analysis of the Ego, one often gets the impression that he was a realist about the Ego. In my opinion this would be a very plausible interpretation of his views, though, this was not what he said in *Experience and Prediction*. However, 13 years later, in *The Rise of Scientific Philosophy* from 1951. he takes a realist stance about our own existence and says:

> We have no absolutely conclusive evidence that there is a physical world and we have no absolutely conclusive evidence either that we exist. But we have good inductive evidence for both assumptions. ... we have good reasons to *posit* the existence of the external world as well as that of our personalities. All our knowledge is posit; so, our most general knowledge, that of the existence of the physical world and of us human beings within it, is a posit. (Reichenbach 1951: 268)

Today we would say that he was a *critical realist* here, or even that he relied on *the inference to the best explanation*: I am justified in believing that I exist because the assumption that I exist is the best explanation of a number of phenomena.[16]

4.4. Reichenbach on *Cogito*

Although a critique of *Cogito* is presented in the previous chapter of the article, Reichenbach's critique will be presented in this chapter because it presupposes his positive views about the nature of the self. No matter whether Reichenbach's conclusion about the ontological status of the self in *Experience and Prediction* is right or not, his analysis of the Descartes' argument is detailed and excellent. Talking about the *Cogito*, he says:

> There is a long line of experience hidden behind this "I." The ego is by no means a directly observed entity; it is an abstractum constructed of concreta and illata as internal elements. Descartes's idea that the ego is the only thing directly know to us and of which we are absolutely sure, is one of the landmarks on the blind alley of traditional philosophy. It involves mistaking an

---

[16] It is plausible to interpret Reichenbach as a realist or rather as a proto-realist. In *Experience and Prediction,* §14 *A cubical world as a model of inferences to unobservable things* he describes, and prescribes, how we should infer the existence of unobservable things.

abstractum for a directly observed entity, mistaking an empirical fact for a priori knowledge, mistaking a product of experience and inferences for the metaphysical basis of the world. Empiricists of all times have rightly opposed it. (Reichenbach 1938: 261)

(Of course, at this point he quotes Lichtenberg.) Reichenbach's critique of Descartes' *Cogito* can be summed up in the following five points:

(1) Self is not something simple, it is something composed of elements.
(2) Self is not known by a direct insight, but indirectly and gradually.
(3) Self is not the Archimedean point of the knowledge, it is discovered later in the process of the rational reconstruction.
(4) Self is not known *a priori* but *a posteriori*, its existence is an empirical discovery.
(5) Self is not something that exists necessarily, its existence is contingent.

In order to fully understand Reichenbach's critique, a crucial thing to have in mind is that he was a *direct realist*. He believed that what we perceive are physical objects in the world, not our impressions. The idea that we have impressions is not an immediately given fact of the consciousness, it is *a result of the epistemological reflection*. For Reichenbach, impressions are *illata*, inferred entities, theoretical entities, not something that is immediately given to us. The consequence of this difference in status is the difference in the level of certainty. For Descartes, when I think that I have the impression of *X I can be absolutely sure* that I really have the impression of *X* because I am immediately aware of it. However, for Reichenbach, when I think that I have the impression of *X I cannot be absolutely sure* that I really have it because theoretical inference can always be wrong. If my theory is wrong, then I do not have the impression of *X*, rather something else is going on. In *Experience and Prediction* Reichenbach dedicates a whole part of the book to impressions, *Part III. AN INQUIRY CONCERNING IMPRESSIONS*, especially §19. *Do we observe impressions?*

What I observe are things, not impressions. I see tables, and houses, and thermometers, and trees, and men, and the sun, and many other things in the crude sphere of crude physical objects; but I have never seen my impressions of these things. ... I believe that there are impressions; but I have never *sensed* them. When I consider this question in an unprejudicated manner, I find that I *infer* the existence of my impressions. ... The distinction between the world of things and the world of impressions or representations is therefore the result of epistemological reflection. (Reichenbach 1938: 162, 163)

Now, let's go back to the above list of the five points.

(1) For Descartes self is something simple. It is a substance (i) to which attributes are attached and (ii) which serves as a principle of individuation.

In the *Sixth Meditation* Descartes says: "When I consider my mind, that is to say myself in so far as I am only a thinking thing, I can distinguish no parts, but conceive myself as one single and complete thing." (Descartes 1641: 164) In opposition to this, Reichenbach holds that self is composed. Abstractum is something that is essentially composed of elements. (Perhaps this explains Reichenbach's choice, why he argued that we are abstracta rather than concreta.) So to say, for Descartes I *am* a simple substance to which different attributes are attached to, while for Reichenbach I just *am* the elements that I am composed of.

(2) It would not be quite correct to say that for Descartes self is known by a direct insight. As we saw at the beginning of this article, in the analysis of *Cogito*, for Descartes self is inferred, not directly given in the experience. Though, not much is needed for this inference. A single thought plus the axiom that a thought cannot exist without the one who thinks it. On the other hand, for Reichenbach the discovery of the self is a whole epistemological process, certainly not a single step. We have to know a lot before we have a right to claim our own existence. Of course, Descartes was not naive. He knew that the discovery of the self is a long process. In the *Sixth Meditation* he talks about this process:

> Firstly then, I perceived that I had a head, hands, feet and all the other members of which body that I considered as a part, or perhaps also as the whole of me, is composed. Further, I perceived that this body was placed among many others, from which it was capable of receiving various agreeable an disagreeable effects, and the agreeable ones I observed by a certain feeling of pleasure, and the disagreeable ones by the feeling of pain. And besides this pleasure and pain, I also felt within me hunger, thirst and other similar appetites, as also certain composed inclinations toward joy, sadness, anger and other similar passions. (Descartes 1641: 152, 153)

One might wonder where is the relevant difference between this description of Descartes and previously quoted description of Reichenbach. Details aside, they both described the same process. So, what's the difference? The difference lies in the fact that, although they both described the same process, for Descartes this description is *explanatory* only while for Reichenbach it is also *justificatory*. And this brings us to the next point.

(3) Descartes and Reichenbach both wanted the same thing, they wanted to justify our beliefs, they wanted to prove that we have knowledge. However, for Descartes the chain of justification starts with *the Cogito*, while for Reichenbach it starts with *the Given*. In other words, they differ in their choice of the Archimedean point of knowledge. For Descartes it is *the Cogito*, for Reichenbach it is *the immediate experience*, that is, *the Given*. Also, for Descartes *Cogito* is indubitable, while for Reichenbach it is not indubitable because it is grounded in the fallible theoretical inference

that I have impressions, not in the infallible immediate awareness of these impressions. Thus in *The Rise of Scientific Philosophy* from 1951, §3. *The Search for Certainty and Rationalistic Conception of Knowledge,* Reichenbach says:

> If the existence of the ego is not warranted by immediate awareness, its existence cannot be asserted with higher certainty than that of other objects derived by means of plausible additions to observational data. (Reichenbach 1951: 35)

Besides, Reichenbach was a fallibilist and was not impressed with rationalistic search for certainty. He even made a Freudian diagnosis of Descartes' search for certainty: "this man needed his philosophical system in order to overcome a deeply rooted complex of uncertainty." (Reichenbach 1951: 36)

(4) Strictly speaking, *Cogito* is not completely *a priori*. Its first part "I think." (or "There is thinking now." or "There is a thought now.") is *a posteriori*. Of course, it is neither a proper empirical knowledge about the facts in the world because it is supposed to be obtained through the introspection about my own mental states. The inference to the "I am" relies on the *a priori* common notion or axiom that one who thinks has to exist in order to think, or on the *a priori* axiom of the S-A ontology that attribute has to be attached to a substance. In this sense, for Descartes the discovery of my own existence is *a priori*. My own existence is a truth of reason, given that there is a single thought. On the other hand, for Reichenbach the discovery of my own existence is completely *a posteriori*. Bodily states are discovered through experience, while impressions or representations are posited in order to explain certain empirical phenomena. "It is the empirical discovery of the difference between the subjective and the objective world which is expressed by the use of 'I'." (Reichenbach 1938: 260) One's own existence is completely empirical fact, even for the one who discovers it from the first person perspective. Reason alone cannot tell me that I exist. Avicenna's *floating man* could not find out that he exists.

(5) For Descartes, as soon as there is experience, there *must* be an I who experiences. That is, given a single piece of experience, my own existence is necessary. However, for Reichenbach, even when there is experience, it is still an open question whether there is an I who experiences. A proper rational reconstruction of experience can, but need not, lead to the discovery of the self whose experience it is. So, even when experience exists, my own existence is still contingent. This is a very strong claim. How could experience exist without somebody whose experience it is? Descartes thought something like this is inconceivable. In the *Sixth Mediatation* he says:

> I find in me faculties of thought altogether special and distinct from myself, such as the faculties of imagination and perceiving, without which I can indeed conceive myself clearly and distinctly as whole and entire, but I cannot conceive them without me, that is to say, without an intelligent substance to which they are attached. (Descartes 1641: 156, 157)

What is implausible here is the claim that even without the faculties of imagination and perception I would still be "whole and entire." What is certainly plausible here is the claim that these faculties cannot exist "without an intelligent substance to which they are attached." However, Reichenbach accepts the challenge and argues that in principle there can be experience without a self to which it belongs:

> As the abstractum "ego" is to express an empirical fact, we are free to imagine a world in which there would be no ego. Imagine that all people were connected, according to the salamander operation (§27), in such a way that everybody shared the impressions of everybody else. Nobody would then say, I see, or I feel; they would all say, There is. On the other hand, we may obtain the opposite case by dissolving the unity of one persona into different egos at different times; if there were no memory, the states of one person at different times would be divided into different persons in the same way that spatially different bodies are divided into different persons. The concept of ego then would not have been developed. (Reichenbach 1938: 261, 262)

It is hard to say how things would look like if 7 billions of us were all connected in such a way. There would be no individual selves but such a grotesque creature could develop some sense of a self, as distinct from mountains and oceans. Probably some errors in perception would occur and be corrected later. In that case a creature could understand a difference between *I see* and *There is*. A creature would probably not develop a concept of heterophyschological, it would be lonely. Perhaps, contrary to Reichenbach's intuitions, it would come to the conclusion *I think, therefore I am*. Though, it is not clear how it could formulate it. A creature would probably not develop a language because it would not need a verbal communication. The opposite case is also not quite clear. People without memory could not learn anything, they could not understand anything. Assume that our memory is being wiped every day at midnight, or 1st of January every year. That would still be a too short period to develop selves. We could say that in a sense there would be selves but they would not last long enough to understand that they exist. No matter how convincing we find these thought experiments of Reichenbach, he did hit at the right place. He did not want to show that in the empty universe there would be no selves. This claim would be trivial. He wanted to show that, even in the universe in which there was experience, there still might be no selves. And this is an interesting and very strong claim.

Reichenbach argues that my own existence is a *hypothesis* for me. But if it is a hypothesis then it must be in principle possible that it is false. And this means that it must in principle be possible that I only think that I exist but that I do not really exist. But how could that be? How could I think that I exist if I do not exist? This is the Descartes' foothold. Here we are not talking about the feeling of authenticity that we might sometimes have. We may say "I did not exist until I discovered my true self." But this is only a metaphor. Here we talk literally about our own existence. I can imagine a scenario where I wake up and discover that everything up until this morning was a dream. I could imagine a scenario in which I discover that I am a brain in a vat in the laboratory somewhere at the Alpha Centaury. In these radical sceptical scenarios I would find out that I had completely wrong beliefs about my own nature and position in the world, but these discoveries would be discoveries about *me*, the same *thinking subject*, the same *subject of experience*. But how could I imagine a scenario that one morning I wake up and discover that I do not exist, and that I never did exist? Who is discovering that if I do not exist? Reichenbach's salamander operation thought experiment describes such a situation. We can imagine that one human body, although in fact a part of collective consciousness, wrongly believes that it has individual existence. One day he discovers that he does not have independent existence but that he is just a part of collective consciousness. In fact, this idea is present in religion and science fiction. Perhaps Buddhist No-Self View is true and we do not really have individual existence. We may wake up one morning with that revelation. In *Star Trek* a Borg drone might have a fever and hallucinate that he exists on his own. As if my little finger hallucinated that it existed on its own but in fact it did not. Odo might immerse himself into the Great Link and end the illusion of independent existence. In a sense, we can understand such scenarios, but still the question remains. What sense does it make to claim that for 50 years I falsely believed that I existed but that in fact I did not exist? Even if tomorrow morning I merge myself into a huge cosmic soul, if for 50 years I believed that I existed, then I existed for 50 years. Even if my beliefs about myself were massively wrong, they were *my* beliefs, beliefs of a single subject of thinking and experience. Real people in *Matrix*, who lay intubated in baths, are still subjects of experience, although they have completely wrong beliefs about themselves.

5) Reductionism and Circularity

Generally speaking, there are three possible views about the self. (1) *Antireductionism* - Self exists on its own. It exists *per se*. It is something that has experience, memory, body, character traits, etc. but in principle it can

exist independently of these elements. (2) *Reductionism* - Self exists. However, it is nothing but its experience, memory, body, character traits, etc. It has no existence over and above the elements that it is composed of. (3) *Eliminativism* - Self does not exist. We only think that there are such things as selves but in reality such things simply do not exist. Logical positivists were reductionists about the self. They believed that the self existed but that it was reducible to experience.

Standard objection to the reductionism about the self is *circularity*.[17] For reductionists the self is usually seen as something that is reducible to experience. However, the problem with this option is that experience is not something that can exist on its own. It can exist only if it belongs to somebody whose experience it is. Talk about experience implicitly presupposes self who has that experience. The concept of experience implies the concept of self. Carnap himself was well aware of this fact. In the *Aufbau* §18 *The Physical and Psychological Objects* he says that "psychological objects have the positive characteristic that each of them belongs to some individual subject." (Carnap 1928: 33) Perhaps we can understand the general spirit of the Lichtenberg's comment that we should say *It thinks!* just as we say *It lightens!* But in its literal meaning, the comment is not clear. Lightening does not need a subject, but thinking does. There can be a lightening without Zeus, Perun, St.Elias, or someone who lightens, but there cannot be thinking without someone who thinks. We cannot take "Love is in the air!" in its literal meaning. It is only a metaphor. So, the objection runs that we cannot define the self in terms of experience because experience presupposes the self. In such a definition an explanans would contain an explanandum. A reductive sentence of the form:

$$X \text{ is nothing but } a, b, c, ...$$

cannot fulfill its reductive purpose because the meaning of "*a*, *b*, *c*, ..." entails that there must be an *X* to which they belong.

Logical positivists were well aware of this problem and they had an elaborated answer to it. The problem, as well as its solution, can be best understood within the framework of the positivistic *constructional systems*. That is, Carnap's *The Logical Strucutre of the World* from 1928 and Reichenbach's *Experience and Prediction* from 1938. On the one hand, they wanted to show how the self is constructed out of the elementary experiences, that is, out of the given. On the other hand, they started their constructional systems with the *elementary experiences*. But whose experiences? As we saw, experience has to be somebody's experience. Does it mean that there

---

is a concealed epistemic subject already at the very beginning of their constructional systems? Does it mean that Carnap and Reichenbach in fact started their epistemic endeavours from their own experience, just as Descartes did? Well, in a sense they did. However, in a relevant sense they did not. §65 of the *Aufbau* has the indicative title *The Given Does Not Have a Subject*. In that paragraph Carnap explains:

> In our system form, the basic elements are to be called experiences of the self after the construction has been carried out; hence, we say: in our constructional system, "my experiences" are the basic elements ... the characterisations of the basic elements of our constructional system as "autopsychological", i.e., as "psychological" and as "mine", becomes meaningful only after the domains of the nonpsychological (to begin with, the physical) and of the "you" have been constructed. (Carnap 1928: 104)

In order to be completely clear about it, and in order to avoid vicious circle, in §75 Carnap draws a distinction between *factual language* and *constructional language*. The expressions of the factual language he marks with the index "P" and the expressions of the constructional language with the index "C". He relies on this distinction already in §64 *The Choice of the Autopsychological Basis*, where he says:

> We prefer to speak of the *stream of experience*. The basis could also be described as *the given*. But we must realize that this does not presuppose somebody or something to whom the given is given. The expression "the given" has the advantage of a certain neutrality over the expression "the autopsychological" and "stream of experience." Strictly speaking, the expression "autopsychological" and "stream of experience" should be written in the symbolism introduced in §75 as ᴾautopsychologicalᴾ and ᴾstream of experienceᴾ. (Carnap 1928: 101, 102)

So, although the basis of the constructional system is ᴾmy own experienceᴾ the justification is not circular because it starts with the ᶜsubjectless givenᶜ. Although I know that ᶜsubjectless givenᶜ is in fact ᴾmy own experienceᴾ, I have to start the process of rational reconstruction from the ᶜgivenᶜ. Does it mean that I have to start the process of justification of all of my beliefs from my own experience and pretend that I do not know that it is my own experience? Well, yes! I can justify all of my beliefs only if I sincerely pretend that I do not know that the starting point is my own experience. It is the only way in which I can justify my beliefs that my own experience is experience and that it is mine. The claim that ᶜIᶜ am constructed out of ᴾmy own experienceᴾ would be circular and uninteresting, but the claim that ᶜIᶜ am constructed out of ᶜmy own experienceᶜ is a valuable theoretical insight into my own nature. And this is the claim of the reductionism about the self: I have to show how am I constructed out of the given, that is, how am I constructed out of elements that do not already contain I. Reduction-

ist in general has to show how *X* comes into existence out of elements that do not contain or presuppose *X*.

In *Experience and Prediction* §28 *What is the Ego?* Reichenbach offers the same answer.[18] He says that he uses the ego-language just for the sake of convenience. He holds that all the facts that lead to the discovery of the ego in principle can be described in a neutral language, without using the concept of the ego.

> We described, some lines previously, the facts leading to the discovery of the ego, and said "*We* stand at the window and see a car … another person … tells *us* …" Thus in this description we already used the ego-language which we wanted to substantiate. This is, however, no contradiction or vicious circle. We used the usual ego-language only to be more easily understood. We could have given the same description by speaking in a neutral language. The original neutral language does not say "I see" but "There is"; only because we hear that another person answers "There is not" do we retire to the more modest statement "I see." (Reichenbach 1938: 260)

Reductionism about the *X* is usually expressed by sentences like "*X* is nothing but *a*, *b*, *c*, …" or "*X* is nothing over and above *a*, *b*, *c*, …," where *a*, *b*, *c*, … are the elements that *X* is composed of. Such reductive sentences can be understood in at least three senses: (1) semantic, (2) epistemological, and (3) ontological. Generally speaking, logical positivists were reductionist about the self in all of these senses. (1) *Semantic reductionism* is the view that when we talk about *X* we in fact talk about *a*, *b*, *c*, … "*X*" does not have any meaning on its own, different from the meaning of "*a*, *b*, *c*, …" This is a semantic reductionistic thesis about the meaning of "*X*." Alternative might be a sort of *error theory* - a claim that "*X*" has a meaning of its own, but, since no corresponding entity exists, it does not refer at all. A thesis of semantic reductionism might be expressed as a claim that *X–language* is in principle replaceable with the *a,b,c,–language*, without a loss of meaning. (3) *Ontological reductionism* is the view that *X* has no existence on its own, besides the existence of its elements *a*, *b*, *c*, … Whenever *a*, *b*, *c*, … are given, *X* is given as well. *X* has no causal powers distinct from the causal powers of *a*, *b*, *c*, … (2) *Epistemological reductionism* is a less frequent view, but it is perhaps the most interesting one in this context. It is the view that we cannot know *X* unless we know *a*, *b*, *c*, … The knowledge of *X* presupposes the knowledge of *a*, *b*, *c*, … Or, the only way that we can know *X* is that we know *a*, *b*, *c*, … Carnap explains the idea in the *Aufbau*, §54. *Epistemic Primacy*. We have to have in mind that the constructional systems of logical positivists were primarily epistemological systems, they were organized in

---

[18] Although, as we saw, Reichenbach accepts an even stronger challenge and tries to show that experience can exist without the subject.

the epistemological order. Although Carnap talks about the construction of concepts and objects, his overall aim is epistemological. He wanted to justify our beliefs. After all, it was Reichenbach, in the preface to the *Experience and Prediction*, who introduced the distinction between the *context of discovery* and the *context of justification*. His interest was justification of our beliefs. So, when we say that logical positivists were reductionists about the self, we have to emphasise that their reductionism was primarily *epistemological*. Their point was that we cannot know what self is before we know what body is, what senses are, what experience is, what other minds are, etc. Lichtenberg's notice that we should say *It thinks!* as we say *It lightens!* makes more sense if it is understood as a notice about the place of the I-beliefs in the overall epistemological order. In fact, *Cogito* can be formulated within the positivistic constructional system. However, it cannot stand at its beginning. The rational reconstruction of our beliefs has to start much earlier.

REFERENCES

Aristotle (1963). *Categories and De Interpretatione*. Oxford: Clarendon Press. 2002.

Ayer, A. J. (1936). *Language, Truth and Logic*. Penguin Books. 1987.

Ayer, A. J. (ed.) (1959). *Logical Positivism*. The Free Press.

Baker Rudder, L. (2013). *Naturalism and the First-Person Perspective*. Oxford University Press.

Berčić, B. (2004). "Carnap's Loop." *Synthesis Philosophica* 19: 297-306.

Berčić, B. (2002). *Filozofija Bečkog kruga (Philosophy of the Vienna Circle)*, Zagreb: KruZak.

Blackburn, S. (1999). *Think*. Oxford University Press.

Blackmore, J. (1972). *Ernst Mach: His Work, Life and Influence*. University of California Press.

Carnap, R. (1928). *The Logical Structure of the World* (*Der Logische Aufbau der Welt*). University of California Press. 1969.

Carnap, R. (1932). "The Elimination of Metaphysics through Logical Analysis of Language." In Ayer A. J. (ed.) *Logical Positivism* (1959). ("Uberwindung der Metapysik durch Logische Analyse der Sprache." *Erkenntnis* Vol. II)

Carnap, R. (1950). "Empiricism, Semantics and Ontology." *Revue International de Philosophie* 4. In Linsky (ed.) 1952.

Descartes, R. (1641). *Meditations*. Penguin Books (1985).

Descartes, R. (1644). *The Principles of Philosophy*. Blackmasks Online. 2002. (http://www.blackmask.com)

Feigl, H. & Sellars, W. (eds.) (1949). *Readings in Philosophical Analysis*. Appleton-Century-Crofts.

Hintikka, J. (1962). "*Cogito, Ergo Sum:* Inference or Performance." *The Philosophical Review* LXXI: 3-32.

Lichtenberg, G. (2012). *Philosophical Writings*. SUNY Press.

Linsky, L. (ed.) (1952). *Semantics and the Philosophy of Language*. The University of Illinois Press et Urbana.

Locke, J. (1690). *An Essay Concerning Human Understanding*. Penguin Books (1997).

Mach, E. (1886). *The Analysis of Sensations* (*Die Analyse der Empfindungen*). The Open Court Publishing Company. 1914.

Nagel, T. (1986). *The View From Nowhere*. Oxford University Press.

Reichenbach, H. (1938). *Experience and Prediction*. The University of Chicago Press (1966).

Reichenbach, H. (1951). *The Rise of Scientific Philosophy*. The University of California Press.

Russell, B. (1914). "The Relation of Sense-data to Physics." In *Mysticism and Logic and Other Essays*. London: Allen & Unwin (1959).

Schlick, M. (1918). *General Theory of Knowledge (Allgemeine Erkenntnislehre)*. Wien, New York: Springer-Verlag 1974. (translation of the 2nd edition from 1925).

Schlick, M. (1934). "The Foundation of Knowledge." In Ayer A. J. (ed.) *Logical Positivism* (1959). ("Uber das Fundament der Erkenntnis." *Erkenntnis* Vol. IV).

Schlick, M. (1936). "Meaning and Verification." *The Philosophical Review* 45, in Feigl & Sellars (eds.) *Readings in Philosophical Analysis* (1949).

Strawson, Peter F. (1959). *Individuals*. Routledge. 1996.

Von Mises, R. (1939). *Kleines Lehrbuch des Positivismus*. Frankfurt am Main: Suhrkamp. 1990.

Weinberg, J. R. (1936.) *An Examination Of Logical Positivism*. London: Kegan Paul.

Williams, B. (1978). *Descartes: The Project of Pure Enquiry*. Penguin Books.

Wittgenstein L. (1953). *Philosophical Investigations*. Basil Blackwell. 1986.

7. Brentano On Self-Consciousness[1]

LJUDEVIT HANŽEK

1. Introduction

Austrian philosopher Franz Brentano (1838-1917) offered a highly elaborated and idiosyncratic theory of consciousness and self-consciousness in his masterpiece, *Psychology from an Empirical Standpoint* (1874). In this paper, I discuss the main and most general points of this theory. While I do refer to Brentano's original work and evaluate his arguments, this paper does not aim for a thorough historical examination of Brentano's work. I am interested in the basic elements of his approach to self-consciousness, and the reasons supporting them.

I begin by providing a minimal background to his ideas, then proceed to clarify his main points on self-consciousness. After that, I analyze the arguments in favor of, and objections to Brentano's theory; I also mention a possible reinterpretation of his theory, which I do not find plausible.

2. Background

Introspection is the cognitive mechanism which gives rise to a subject's awareness of his own mental states. In the philosophical tradition, it is often considered to be a form of inner observation, in contrast to perception (external observation), the cognitive mechanism which results in the awareness of the external world. Brentano's thoughts on self-consciousness are closely related to the issue of introspection, and the problems surrounding the topic in late 19th century. It is important to bear in mind that the philosophical views on introspection at the time differed significantly between the German philosophical tradition and the British one.

Danziger lists the following reasons as a ground for skepticism about introspection in the German tradition: first, the influence of Kant – who maintained that all the information that can be attained by the subject's

observation of his mental states relates only to the phenomenal self, not the real self (Danziger 1980: 242); secondly, the influence of German idealism (particularly the Hegelian school), according to which introspection reveals only the individual self, while the true Self is revealed in society, culture and history (Danziger 1980: 243); finally, the strong influence of Lange, who attacked introspection as a method of psychological research, and claimed that the psychological activity should be studied in terms of its material manifestations (Danziger 1980: 243).

On the other hand, the British philosophical tradition, in accordance with its thoroughgoing empiricism, stressed the importance of observation of one's own mental states, believing it to be the only way to acquire evidence for claims about the structure of the mind. (Danziger 1980: 242)

3. Brentano on Inner Consciousness

Despite being a part of the German speaking cultural circle, Brentano's philosophical outlook in *Psychology from an Empirical Standpoint* is, in concordance with the title of the book, decidedly empiricist. The central idea of the book is to establish a psychology based both on inner and outer experience. (Brentano 1995a: 22) In his book, Brentano refers to many authors from the history of philosophy as well as his contemporaries, and several of them belong to the British empiricist tradition; a particularly prominent one is John Stuart Mill.

Nevertheless, in contrast to the prevailing attitude in the empiricist camp, Brentano accepts the objections to the possibility of introspection that were circulating in the philosophical literature in the second half of the nineteen century. He refers to the views of Lange, Maudsley and Comte, with an emphasis on the latter's argument: introspection, or inner observation, would require an impossible division – it would require the division of the subject into both the observer and the observed at the same time, and that is not possible.[2] Brentano himself argues that introspection

---

[2] "As for observing in the same manner intellectual phenomena while they are taking place, this is clearly impossible. The thinking subject cannot divide himself into two parts, one of which would reason, while the other would observe its reasoning. In this instance, the observing and the observed organ being identical, how could observation take place? The very principle upon which this so-called psychological method is based, therefore, is invalid. Moreover, let us consider to what entirely contradictory procedures this method immediately leads. On the one hand, we are told to isolate ourselves as much as possible from every external sensation, and especially to restrain ourselves from all intellectual work; even if we were only dealing with the most simple mathematical calculation, what would then happen to "inner" observation? On the other hand, after having finally attained through these measures this state of perfect intellectual sleep, we should devote ourselves to the contemplation of the operations which are

is impossible because any attempt to observe a mental phenomena within the subject himself inevitably leads to the distortion and eventual destruction of the said phenomena. (Brentano 1995a: 22)

However, a crucial distinction between Brentano and the previously mentioned authors is that Brentano distinguishes two different forms of awareness of one's own mental states, of which inner observation (i.e. introspection), is only one. Introspection, as the subject's observation of his current mental states is impossible; but *inner consciousness*, another form of self-awareness, is not only possible, it is the source of all knowledge that the subject can have about his mental states. (Brentano 1995a: 22).[3]

The direction of the subject's attention grounds the difference between inner observation and inner consciousness. Observation is a type of a mental act in which one is focused on the object of the mental act, the object is in the center of the attention. Applied to mental states as objects, this would mean that one would have to direct his attention at his current mental states, and Brentano thinks that this is not possible. However, inner consciousness is an incidental awareness of one's own mental states, contemporaneous with the observation of another object. While the subject is aware of some object, he is also peripherally aware, on the side, of that very mental state.[4] No multiplication of the subject into the observer and the observed is necessary; there is only one mental state, which observes a phenomenon, and which is peripherally aware of itself by means of inner consciousness. The object at which the subject's attention is focused is the primary object of the mental state, while the mental state itself is its secondary object (due to that, Brentano often refers to inner consciousness as secondary consciousness, in contrast to primary consciousness, which is awareness of the primary object).[5]

---

occurring in our mind when nothing goes on in it any longer. To their amusement, our descendants will undoubtedly witness the disavowal of such an assumption." (Comte, as quoted in Brentano 1995a: 24)

[3] Brentano refers to inner consciousness also as *inner perception*, and as *secondary consciousness*. The terms are synonymous, and only stylistic reasons drive the choice of the expression in a particular context. I will treat them the same way in this paper.

[4] Brentano speaks indiscriminately of both the subject being aware/conscious of a mental state, and a mental state being aware/conscious of a mental state.

[5] "One observation is supposed to be capable of being directed upon another observation, but not upon itself. The truth is that something which is only the secondary object of an act can undoubtedly be an object of consciousness in this act, but cannot be an object of observation in it. Observation requires that one turn his attention to an object as a primary object. Consequently, an act existing within us could only be observed by means of a second, simultaneous act directed toward it as its primary object. There just is no such accompanying inner presentation of a second act, however. Thus we see that no simultaneous observation of one's own act of observation or of any other of one's

In a paper, which presents a broadly Brentanian position, Uriah Kriegel invokes the distinction between focal awareness and peripheral awareness and compares it to that between foveal vision and peripheral vision. (Kriegel 2004: 190) While a computer screen is in the center of my visual field (the object of foveal vision), there are objects I am visually aware of that are not in the center of my visual field, e.g. a telephone (the object of peripheral vision). However, the fact that an object is not in the center of my visual field does not mean I am not visually aware of it, for there is a clear distinction between that object (e.g. a telephone) and the objects that are not a part of my visual field at all, e.g. a car on a remote parking lot. (Kriegel 2004: 190) This example from the field of vision can be translated into a more general distinction between focal awareness of an object – the object being at the center of attention – and peripheral awareness, where the object is not at the center of attention. While not discussing Brentano's views explicitly, Kriegel's distinction between focal and peripheral awareness closely corresponds to Brentano's distinction between primary consciousness and secondary consciousness, which means that the analogy with foveal and peripheral vision also fits well with Brentano's position.

Brentano puts forth his theory in the context of debating the Regress Argument for the existence of unconscious mental states. (Brentano 1995a: 93-94) The Regress Argument purports to establish the existence of unconscious mental states as the only scenario which avoids the existence of an infinite number of mental states.[6]

The Regress Argument:
1. For every *conscious* mental state **M**, there is a higher-order mental state, **M***, such that **M*** is *conscious of* (*aware of*/*represents*) **M**.

---

own mental acts is possible at all. We can observe the sounds we hear, but we cannot observe our hearing of the sounds, for the hearing itself is only apprehended concomitantly in the hearing of sounds." (Brentano 1995a: 99)
From Brentano's explanation of the difference between primary and secondary consciousness, it is clear that bridging the ontological gap between the subject and the object of the observation is very important to him:
"The presentation of the sound and the presentation of the presentation of the sound form a single mental phenomenon; it is only by considering it in its relation to two different objects, one of which is a physical phenomenon and the other a mental phenomenon, that we divide it conceptually into two presentations. In the same mental phenomenon in which the sound is present to our minds we simultaneously apprehend the mental phenomenon itself. What is more, we apprehend it in accordance with its dual nature insofar as it has the sound as content within it, and insofar as it has itself as content at the same time." (Brentano 1995a: 98)
[6] Brentano refers to discussions of the Regress Argument in Herbart and Aristotle (Brentano 1995a: 94). The modern version of the Regress Argument is in Rosenthal (Rosenthal 1986: 340).

2. If all mental states are conscious, then there is an infinite number of higher-order mental states; otherwise, there would be such a mental state **M****, for which there would be no higher-order state **M*****, which is *conscious of* **M**** - which means that **M**** would be unconscious.

3. It is absurd to postulate an infinite number of higher-order mental states.

4. Therefore, there is a mental state **M****, for which there is no higher-order state **M*****, which is *conscious of* **M****; i.e. there are unconscious mental states.[7]

The premise that does most of the theoretical work is obviously (1). It links the property of a mental state's being *conscious* with the property of a numerically distinct mental state being *conscious of* the former.[8] What makes a mental state conscious is simply the fact that there is another mental state, which is aware of or represents the first mental state – consciousness is defined in terms of *meta-representation* (specifically, *higher-order representation*). Premise (2) is a trivial corollary, while (3) is a plausible claim, given that the complexity of the human cognitive system and its neurophysiological underpinnings seems to be finite, even if it is extraordinary.[9]

The relation of *consciousness* with *consciousness of* is a vexing problem in contemporary philosophy of mind. There it is usually described in terms of the distinction between *intransitive* consciousness (*consciousness*) and *transitive* consciousness (*consciousness of*). (Gennaro 2004: 2-3; Kriegel 2004: 182-184) Intransitive consciousness is the property of a mental state to be conscious, which usually amounts to having a phenomenal character. (Gennaro 2004: 2; Byrne 1997: 105; Rosenthal 1986: 351-352) Transi-

---

[7] This is a reconstruction of the argument that Brentano reports. (Brentano 1995a: 94)

[8] There is no circularity in the claim, because the predicate *conscious* is different from the predicate *conscious of*. See footnote 9.

[9] Of course, there is a virtually infinite number of ways in which we could describe human neurophysiological states, but that does not entail the existence of an infinite number of those states. First, the vast majority of the possible descriptions would be unnatural and completely irrelevant in terms of explantory potential ("neurophysiological state whose electrical activity pattern is isomorphic to physical phenomenon X," in which X bears no significant relation to the neurophysiological state); secondly, the fact that a given state is described in different ways does not mean that the number of distinct states is increased; thirdly, even if it turns out that the physical structure of the world is infinitely complex (e.g. there is an infinite number of subatomic levels, composed of ever smaller particles) it would not mean that such infinity yields an infinite number of mental states – the number of distinct physical states comprised by a single neurophysiological states would be infinite, but only the physical states at the appropriate level of complexity would constitute neurophysiological states. Since cognitive complexity is plausibly constrained by neurophysiological complexity, the number of different cognitive states will also be finite.

tive consciousness (*consciousness of*) is a property of a subject to be aware of, or represent something, whether a physical object or a mental state. (Rosenthal 1997: 737) While Rosenthal reserves the property of transitive consciousness for cognitive subjects, Kriegel believes one can plausibly talk about transitive consciousness of mental states: a mental state is transitively conscious of an object X iff it is the case that in virtue of being in that state, the subject is transitively conscious of X. (Kriegel 2004: 183-184) Higher-order theories of consciousness aim to analyze intransitive consciousness in terms of transitive consciousness: a mental state is intransitively conscious iff there is a numerically distinct mental state transitively conscious of it. The details of the analysis vary from one author to another, as there are competing accounts of the exact nature of the relationship between the mental states. However, a unifying thought is the idea that a mental state is phenomenally conscious iff it is represented by a distinct mental state; thus, phenomenal consciousness is reduced to access consciousness of the appropriate type.[10]

While the nineteen century debate on unconscious states lacks the conceptual sophistication of contemporary philosophical work on consciousness, premise (1) in the Regress Argument states a claim essentially identical to modern higher-order views – a mental state is conscious iff another mental state is conscious of it.

Brentano rejects that premise. According to him, postulating the existence of unconscious mental states is not the only solution which avoids the absurdity of an infinite number of mental states. He proposes that, contrary to premise (1), it is not the case that a mental state is conscious only if a numerically distinct mental state is conscious of it; a mental state is conscious if it is conscious of itself. Consciousness still boils down to meta-representation, but not by a numerically distinct state; a mental state's reflexive consciousness of itself, or *self-representation*, is what makes it conscious in the intransitive sense.[11] This consciousness of itself exhibited by a mental state, comes in both the form of *presentation* of the state and in the form of *judgement* affirming the existence of the state. (Brentano 1995a: 119) Presentation is the most generic form of representation for Brentano,

---

[10] This fact clearly shows that the account is indeed not circular, for one type of consciousness is explained by reference to a different type of consciousness. The author of the distinction between phenomenal consciousness and access consciousness is Ned Block. (Block 1995)

[11] "Every mental act is conscious; it includes within it a consciousness of itself. Therefore, every mental act, no matter how simple, has a double object, a primary and a secondary object. The simplest act, for example the act of hearing, has as its primary object the sound, and for its secondary object, itself, the mental phenomenon in which the sound is heard." (Brentano 1995a: 119)

and it is not particularly clear how to understand it – it is perhaps like a concept (or the tokening a concept in the subject's mind). Judgement is the Brentanian equivalent of thought, or belief, with some important differences, which will soon become evident.

This reflexive consciousness is a form of self-consciousness, but it is radically different from the self-consciousness in the form of introspection, and Brentano is very critical of the authors who failed to see the distinction between the two. (Brentano 1995a: 23-24) Whereas introspection does not exist at all, this form of self-consciousness is a fundamental aspect of the mental life, and the foundation of scientific psychology. (Brentano 1995a: 22)[12]

4. Arguments and Criticisms

There are three arguments in favor of this conception of self-consciousness that can be distilled from Brentano's text.

Phenomenological Argument. Brentano in several places states that the construal of self-consciousness as an inner consciousness is an evident fact, a basic datum of experience. Experience clearly shows that the awareness of our own mental states is not in the form of observation, but is incidental and peripheral. (Brentano 1995a: 22, 97-100, 110)

The first problem with the Phenomenological Argument is that an opponent might simply deny the claim and say that there is nothing like Brentano's proposed inner consciousness going on inside one's mind during a normal experience. Appeals to phenomenology in general seem to be a weak type of argument, because they lead to a stalemate when faced with a critic who claims to not share the same phenomenological impression. However, if one could convincingly explain away the competing phenomenology, it might provide the original argument with some justification. Kriegel attempts to do that, while defending his broadly Brentanian theory of consciousness. Kriegel also claims, like Brentano, that there is a form of peripheral self-consciousness, which can be detected in experience. As for criticisms in the form of one failing to detect that phenomenon in their experience, Kriegel responds: first, there is a peripheral awareness of every other sort, so it is only expected that there is a peripheral awareness in the domain of self-consciousness, too. (Kriegel 2004: 191) By peripheral awareness of every other sort, Kriegel refers to perceptual peripheral

---

[12] Brentano's complete theory of consciousness, which centers around intentionality and inner consciousness, is highly complex and elaborate; for the purposes of this paper, it is unnecessary to delve deeper in it.

awareness, intellectual peripheral awareness, and perceptual peripheral awareness simultaneous to an episode of focal intellectual awareness. (Kriegel 2004: 190-191)[13] Secondly, self-consciousness is just not particularly phenomenologically impressive. (Kriegel 2004: 193) For example, Kriegel notes that a denial of this kind of self-consciousness appears much more acceptable than a denial of color experiences would – the reason is the vividness of a color experience, compared to peripheral self-consciousness. (Kriegel 2004: 193) Kriegel thinks that the disagreement on the phenomenology of peripheral self-consciousness is more like the debate about phenomenology of propositional attitudes. Finally, peripheral self-consciousness is pervasive and ubiquitous, it is a constant feature of a human experience. (Kriegel 2004: 194) As such, it is impossible to notice a contrast between the experiences that contain it and those that lack it, just like it is difficult to notice the hum of a refrigerator pump after it has been operating for some period. (Kriegel 2004: 194) Given that Brentano's and Kriegel's theory are almost identical in their general form, all these responses might conceivably be used against a critic of Brentano's claim that the inner consciousness model is self-evident. However, Kriegel's responses do not seem particularly convincing. His reasoning that a peripheral form of self-consciousness is only natural, since there is a peripheral awareness of every other sort, is contentious. Firstly, this kind of analogical reasoning about seemingly contingent matters is not reliable. Secondly, there is a significant distinction between self-consciousness and every other form of awareness mentioned by him. Self-consciousness is the only reflexive form of awareness. Finally, Kriegel's classification of these other forms of awareness is very coarse-grained – he does not distinguish types of intellectual awareness (e.g. believing and hoping), although he does distinguish some types of perceptual awareness (visual and auditory modalities) – and one might worry that the classes are arbitrarily identified. His claim that a critic would fail to notice peripheral self-consciousness because of its weak phenomenological impression is also problematic; a critic might point out that nonexistent things typically are not phenomenologically impressive, and that it is at least as likely that the supposed subtle phenomenology that the proponent of such a phenomenon reports is a cognitive illusion of some sort.

---

[13] The example with foveal and peripheral vision illustrates perceptual peripheral awareness. Intellectual peripheral awareness refers to being aware of a thought on the side, incidentally, as the subject is focused on another thought or inference. Perceptual peripheral awareness during an episode of focal intellectual awareness refers to the subject's peripheral awareness of perceptual stimuli while preoccupied with a thought or an inference.

Another, and a much more serious problem with the Phenomenological Argument is the shift of attention which it induces. By treating the phenomenology of normal experience as the evidence for his model of self-consciousness, Brentano turns his attention towards that phenomenology, i.e., he focuses on his mental states. But according to him, that procedure would by that feature become inner observation, and he claims that inner observation is impossible. The objection here is not that there is no such thing as inner consciousness of the sort Brentano espouses – the objection is that any attempt to use one's phenomenology as the evidence for that proposition would be self-defeating. Referring to one's own experience as evidence during an argument, or a premise during an inference, would result in one's attending to their experience – for one could not know that their experience is such-and-such, without examining it; but attending to the experience would turn the attention away from the objects of the experience and to the mental states that constitute the experience, and that would turn experience into inner observation (introspection), which is impossible according to the theory under consideration. Therefore, either Brentano is wrong about the impossibility of inner observation, or he cannot employ the phenomenology of normal experience as evidence for his theory.

It is not clear whether the problem can be mitigated by referring to one's memory of an experience in their immediate past.[14] Brentano thinks that memory is a form of inner observation, but that it is unreliable, introduces the possibility of self-deception and cannot be considered a credible source of evidence about one's mental life. (Brentano 1995a: 26) At the same time, he does acknowledge the fact that, without memory, no experimental science would be possible at all, and he seems to think that memory does not interfere with the past mental phenomena in the same way that a current observation of a phenomena would. (Brentano 1995a: 26) It seems, based on these somewhat conflicting remarks, that Brentano's position on the role of memory in the Phenomenological Argument is not clear.[15] Thus, the argument remains contentious, at best.[16]

---

[14] Sometimes the requirement that introspection must focus on one's current mental status is weakened, by requiring that it focuses on past (or future) mental states within a very short temporal distance. (Schwitzgebel 2016)

[15] I am grateful to Dario Škarica, whose remarks helped me in arriving at a more nuanced reading of Brentano's position on this issue.

[16] A complication arises here. Namely, in his other works, specifically *Descriptive Psychology*, Brentano introduces a form of inner consciousness which seems to function precisely as a mechanism of detecting the features of one's own experience; he calls it "noticing." (Brentano 1995b: 34-66) It is the explicit perception of what was implicitly contained in the subject's consciousness; it is indeed a type of inner consciousness (or

Epistemological argument. Brentano holds that inner consciousness is infallible – it cannot be the case that the subject is aware by means of inner consciousness of a mental state within himself, without that mental state existing within the subject. The infallibility itself is immediately evident, and any attempt of arguing in favor of it is misguided (Brentano 1995a: 109); what is needed is simply an ontological description of the relationship between the elements of inner consciousness which renders it infallible. Furthermore, the infallibility of inner consciousness serves as a ground for any kind of knowledge at all, and doubting it would result in global skepticism.[17]

In discussing the Epistemological Argument, Brentano explains that his model of inner consciousness can explain the infallibility of self-consciousness because it eliminates the ontological distinction between the first-order state and the state which is conscious of it. (Brentano 1995a: 107) If the states in question were numerically distinct, and related by a mere causal relationship, the infallibility of self-consciousness would be impossible. However, the entire argument is somewhat unclear. Firstly, it is not clear what property is Brentano referring to when he says that the infallibility of secondary consciousness is evident – is it evident in the sense in which conceptual truths are evident, or is it evident in the sense in which it is evident that one is undergoing such-and-such experience (the evidence of secondary consciousness itself)? Secondly, Brentano does not clarify the relationship between inner consciousness and other forms of knowledge – if inner consciousness is infallible, and if it is a ground for other forms of knowledge, how is it the case that knowledge about the ex-

---

inner perception), and it is a source of justification. However, a more detailed description of noticing raises doubt about the adequacy of classifying it as inner consciousness – for the purposes of using noticing as method in psychological research, a fitting subject should be chosen, the subject should go through preparations for noticing, and the very act of noticing should be assisted with the instructions for efficient comparisons of different experiences. (Brentano 1995b: 40-52) For example, the subject can be assisted in noticing a red tinge in a blue color by being showed an example of pure blue and then asked to perform a comparison of that example with the example of blue with a red tinge. (Brentano 1995b: 52) It is obvious that the whole procedure described by Brentano would result in the subject's focusing his attention on the colors, which would make it inner observation, not inner consciousness.

[17] "The truth of inner perception cannot be proved in any way. But it has something more than proof; it is immediately evident. If anyone were to mount a skeptical attack against this ultimate foundation of cognition, he would find no other foundation upon which to erect an edifice of knowledge. Thus, there is no need to justify our confidence in inner perception. What is clearly needed instead is a theory about the relation between such perception and its object, which is compatible with its immediate evidence." (Brentano 1995a: 109)

ternal world still does not seem attainable, due to unreliability of external perception? (Brentano 1995a: 108) It seems that the edifice of knowledge built on the infallibility of inner consciousness does not reach the external world, which means that the threat of skepticism is not satisfactorily dealt with, contrary to Brentano's claim.

> **No Unconscious Mental States Argument**. Brentano agrees that, if the premise (1) of the Regress Argument were true, there would be no way of blocking the conclusion that unconscious mental states exist. Since he believes there are no such things as unconscious mental states, he is forced to deny the premise. An account of self-consciousness as inner consciousness is the only possible way of consistently combining the following claims:
>
> 5. Mental state **M** is conscious iff there is a mental state that is conscious of **M**.[18]
> 6. There are no unconscious mental states.
> 7. There is not an infinite number of mental states.

Thus, a strong suspicion towards the idea of unconscious mental states is among the reasons for accepting the inner consciousness account of self-consciousness. (Brentano 1995a: 94)

Brentano's reasons for thinking that all mental states are conscious are numerous and involve a tedious analysis of standardly used arguments in favor of unconscious mental states. (Brentano 1995a: 79-106) He discusses a series of examples which, according to his contemporaries, are best explained by the hypothesis that unconscious mental states exist, and finds them all wanting. The previously mentioned Regress Argument is a *reductio ad absurdum* type of argument in favor of the conclusion that there are unconscious mental states, and it is clear that Brentano's prior suspicion of unconscious mental states functions as one of the reasons for rejecting the Regress Argument. The analysis of Brentano's extremely detailed objections to the arguments in favor of unconscious mental states would be outside the scope of this paper, but what can be said is this: regardless of the fact whether Brentano's views on the matter were justified in his time, the reasons that are today offered in favor of the existence of unconscious mental states are very strong. There are many phenomena discussed in psychology and cognitive science, which can be plausibly explained only under the supposition that there are unconscious mental states affecting the subject's conscious cognition and behavior. As examples, I can mention blindsight and prosopagnosia. *Blindsight* is a condition present in people

---

[18] Crucially, the requirement that **M** and the state that is conscious of **M** be numerically distinct, is dropped.

with damage to the primary visual cortex. These subjects have been shown to detect and discriminate the size and shape of objects present in the parts of their visual field which are inaccessible to their conscious vision. (Augusto 2010) *Prosopagnosia* is a cognitive deficit in virtue of which people are unable to consciously recognize the faces of familiar persons; however, it has been shown that covert facial recognition can take place in people with prosopagnosia, based on the changes in the electrical conductance of the skin. (Augusto 2010) Therefore, even if Brentano was justified in believing that there are no unconscious mental states, today it is almost certain that that belief is false; that means that the No Unconscious Mental States Argument for his model of self-consciousness is not sound. There is nothing problematic with the implication of the Regress Argument that unconscious mental states exist.

Apart from the problems with Brentano's arguments in favor of his model, there are some independent objections.

First, both Brentano and the proponents of the Regress Argument share the belief that a mental state is conscious iff there is a mental state that is conscious of it (meta-representation). Brentano nowhere seems to question that belief, but some contemporary authors do question it. According to representational theories, a mental state is conscious not because it is an object of awareness of some mental state, but because it has an object of awareness – not because it is represented by a mental state, but because it represents something (in the appropriate way).[19] The point of this objection is not that Brentano should have discussed the representational view, the point is rather that the lack of discussion of any kind of alternative to the meta-representational view appears hasty.

Secondly, while Brentano's claim that the mental state represents itself avoids the Regress Problem, it raises concerns over the possibility of another type of regress, which would result in an infinitely complex content of the self-representing mental state.

---

[19] "An experience of x is conscious, not because one is aware of the experience, or aware that one is having it, but because, being a certain sort of representation, it makes one aware of the properties (of x) and objects (x itself) of which it is a (sensory) representation. My visual experience of a barn is conscious, not because I am introspectively aware of it (or introspectively aware that I am having it), but because it (when brought about in the right way) makes me aware of the barn. It enables me to perceive the barn. For the same reason, a certain belief is conscious, not because the believer is conscious of it (or conscious of having it), but because it is a representation that makes one conscious of the fact (that P) that it is a belief about. Experiences and beliefs are conscious, not because you are conscious of them, but because, so to speak, you are conscious *with* them." (Dretske 1993: 280-281)

Internal Regress Problem:

8. If the secondary consciousness of a mental state reduces to the mental state being conscious of itself, it means that the state "encompasses" itself in its totality, i.e. it represents every fact about itself.
9. If secondary consciousness represents every fact about itself, then it must also represent the fact that the mental state is conscious of itself through secondary consciousness.
10. If (9) holds, then there is an additional fact that the mental state must be conscious of – the fact that the mental state is conscious of it being conscious of itself through secondary consciousness.
11. If the mental state is conscious of (10), then it must be conscious of the fact that it is conscious of (10), and the regress into infinitely many facts about the mental state that it must be conscious of, is guaranteed.[20]

The key premise is (8), namely, the requirement that the mental state must represent every fact about itself. If that extends to representational facts about the state, then the regress begins; if not, the Internal Regress Problem is not a decisive objection to Brentano's theory.

What should be distinguished is the representation of *a* mental state that represents itself, from the representation of that mental state *as* the state that represents itself. Representing a self-representing state does not mean that it is represented as a self-representing state; representing it as a self-representing state is a description of a mental state, in which a certain property (self-representation) is predicated to the state. As was already said, Brentano thinks that a mental state represents itself through a judgement, which is a rough (but not perfect) equivalent to the contemporary notion of belief; it seems that that would entail that the mental state contains a descriptive representation of itself. If that it is the case, the Internal Regress Problem seems unavoidable, because a description of itself that would not include its property of self-representation could hardly be called encompassing.[21] However, Brentano's theory of judgement is highly idiosyncratic, and he thinks that judgements do not have to contain predication of a property to the subject, which is particularly true for existential judgements; furthermore, he is explicit about the fact that the judgement

---

[20] This is a reconstruction of the argument found in Zahavi (Zahavi 2006: 3); he reports the argument from Gurwitsch. (Gurwitsch 1979: 89-90) I would like to thank Goran Kardaš, whose question led me to include a discussion of this problem in my paper.

[21] Naturally, one could say that the facts about itself that the mental state should represent exclude representational facts, including self-representational ones. That seems rather arbitrary, though, and conflicts with Brentano's statements. (Brentano 1995a: 98)

of inner perception is not predicative, which means that it is not a description.[22] Essentially, his ability to avoid the Internal Regress Problem depends on the viability of his theory of judgement.

The key part of Brentano's theory of judgement is the claim that the majority of expressions which involve predication of a property to a subject can be reduced to existential expressions, claiming the existence or non-existence of an object; existential expressions are psychologically realized as simple *affirmations* or *denials* of an object, and involve no predication at all. (Brentano 1995a: 150-183)[23] The only thing that is part of the semantic content of an expression is the object itself; the existential import is part of the attitude that the subject takes toward that object – the attitude of affirmation (or *acceptance*) results in a judgement that a certain object exists, while the attitude of denial (or *rejection*) results in a judgement that a certain object does not exist. The issues surrounding Brentano's theory of judgement are numerous and complex, and are not particularly relevant for this paper. However, there is reason to believe that, even on Brentano's theory of judgement, the Internal Regress Problem arises.

Brentano clearly states that the object toward which a subject can take the attitudes of affirmation or denial can be complex – one of the examples he gives is a "sick man" (Brentano 1995a: 165-166); his idea is that the standard view according to which a phrase like "Some man is sick" predicates the property of being sick to the object (man) can be actually be explained away as a simple affirmation of an object, a sick man. Whether or not a combination of concepts should be considered a predication, it is obvious that it is some kind of a description – combining the attribute "sick" with the object "man" results in an object which is different from simply "man." Thus, a description is present in affirmations or denials of complex objects. But that is enough to start the Internal Regress Problem – premise (8) states that the secondary consciousness of a state should represent every fact about the state; the mechanism of that representation is completely irrelevant. The fact that the facts about the state would have to be represented in descriptions like "sick man," instead of explicit predications, does not stop the regress to an infinitely complex description of a mental

---

[22] "No one who pays attention to what goes on within himself when he hears or sees and perceives his act of hearing or seeing could be mistaken about the fact that this judgement of inner perception does not consist in the connection of a mental act as subject with existence as predicate, but consists rather in the simple affirmation of the mental phenomenon which is present in inner consciousness." (Brentano 1995a: 110)

[23] The crucial point here is that there is no combination of a subject with something else, whether that be a predicate or an operator (such as an existential quantifier).

state through its secondary consciousness of itself.[24] Therefore, Brentano is wrong – the Internal Regress Problem cannot be stopped by his theory of judgement.[25]

5. The Adverbial Interpretation

Amie Thomasson offers an alternative interpretation of Brentano's views. (Thomasson 2000) She introduces it in a paper in which she discusses the topic of phenomenal consciousness, and particularly the debate between higher-order theories and representational theories. Thomasson claims that phenomenal consciousness is irreducible, and that Brentano's model of secondary consciousness provides a good illustration of that.

According to Thomasson, it is not necessary to construe Brentano's inner consciousness as consciousness of, or a representation. Inner consciousness is best described as a *way* of being conscious, a *manner* or *mode* of consciousness of some non-mental entity. (Thomasson 2000: 203-204) What makes the difference between conscious and unconscious states is the fact that in conscious states, information is being represented consciously, while the unconscious states represent information unconsciously. Thus, the intrinsic relationship that Brentano claims exists between a mental state and the secondary consciousness of that state actually refers to the fact that representing consciously/unconsciously, or being aware of something consciously/unconsciously, is an intrinsic and constitutive aspect of that mental state, and not something external to it. (Thomasson 2000: 202-204) A mental state is not conscious because some mental state is conscious of it; a mental state being conscious is an irreducible feature of that state, and it is best analyzed as an adverb modifying the awareness of an object that the mental state possesses.

Thomasson's interpretation is interesting, but ultimately unsupported by textual evidence. Brentano regularly speaks of secondary consciousness as consciousness of, and his claim that secondary consciousness contains

---

[24] It is obvious that phrases like "sick man" are actually implicit predications.

[25] Theoretically, Brentano could say that the infinite number of facts about the mental state that the Internal Regress Problem points to can be represented by a description of a sort "mental state representing itself on an infinite number of higher-order levels." That does not sound particularly plausible, however; Brentano insists that the inner consciousness model is evident from experience (Phenomenological Argument), which would mean that this infinite self-representation would also have to be something that the subjects are aware of. He never mentions anything like it, probably because it would be a very implausible suggestion. Secondly, he is clear that the inner consciousness of a mental state is apprehended "in accordance with its dual nature" (Brentano 1995a: 98) – it is doubtful whether the fact of infinite self-representation could be adequately apprehended at all by subjects with finite cognitive capacities.

a judgement further attests that it cannot be interpreted as a purely phenomenal way of awareness. (Brentano 1995a: 100, 107) Also, Thomasson's interpretation completely removes the self-representing aspect from Brentano's theory, even though he clearly states that, in inner consciousness, mental state is aware of itself. (Brentano 1995a: 70, 94, 98)[26] Therefore, Brentano's secondary consciousness is a form of transitive consciousness, and it is distinguished from primary consciousness by the difference in the focus of attention, not by lacking an object.

6. Conclusion

Brentano's view on self-consciousness depends on the distinction between observing one's own mental states and being aware of them only incidentally, at the periphery of consciousness. As I have tried to show in this paper, the arguments he provides in favor of his view are rather controversial, and there are significant difficulties that arise from his claims. It is probable that the central argument for any kind of Brentanian theory of self-consciousness will ultimately have to refer to phenomenology of experience, and as such, it is unlikely that the controversy surrounding his proposal will abate.

REFERENCES

Augusto, L. M. (2010). "Unconscious Knowledge: A Survey." *Advances in Cognitive Psychology* 6: 116-141.

Block, N. (1995). "On a Confusion About a Function of Consciousness." *Brain and Behavioral Sciences* 18 (2): 227-247.

Brentano, F. (1995a). *Psychology from an Empirical Standpoint*. Routledge. German original: Brentano, F. (1874) *Psychologie vom empirischen Standpunkt*. Leipzig: Duncker & Humblot.

Brentano, F. (1995b). *Descriptive Psychology*. London: Routledge. German original: Brentano, Franz (1982) *Deskriptive Psychologie* (ed.) R. Chisholm & W. Baumgartner. Hamburg: Meiner.

---

[26] Thomasson tries to bring the self-representational aspect back by claiming that conscious mental states, though not regularly introspected, are regularly available for introspection, a form of transitive (self)-consciousness (Thomasson 2000: 205). That however, does not eliminate problems, because it makes her theory virtually indistinguishable from higher-order dispositional theories, according to which a mental state is phenomenally conscious if it is available for higher-order representation (Carruthers 2000).

Byrne, A. (1997). "Some Like It HOT: Consciousness and Higher-Order Thoughts." *Philosophical Studies* 2 (2): 103-29.

Carruthers, P. (2000). *Phenomenal Consciousness: a Naturalistic Theory.* Cambridge University Press.

Danziger, K. (1980). "History of Introspection Reconsidered." *Journal of the History of Behavioral Sciences*: 241-262.

Dretske, F. (1993). "Conscious Experience." *Mind* 102: 263–283.

Gennaro, R. J. (2004). "Higher-order theories of consciousness: An overview." In: Gennaro, R. (ed.) *Higher-Order Theories of Consciousness: An Anthology.* Amsterdam/Philadelphia: John Benjamins Publishing Company: 1-13.

Gurwitsch, A. (1979*). Human Encounters in the Social World.* Pittsburgh: Duquesne University Press.

Kriegel, U. (2004). "Consciousness and Self-Consciousness." *The Monist* 87 (2): 182-205.

Rosenthal, D. M. (1986). "Two Concepts of Consciousness." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 49 (3): 329-359.

Rosenthal, D. M. (1997). "A Theory of Consciousness." In: N. Block, O. Flanagan, G. Guzeldere (eds.) *The Nature of Consciousness: Philosophical Debates.* MIT Press.

Schwitzgebel, E. (2016). "Introspection." *Stanford Encyclopedia of Philosophy.* (URL: http://plato.stanford.edu/entries/introspection/ accessed: 22 July 2016)

Thomasson, A. (2000). "After Brentano: A One-Level Theory of Consciousness." *European Journal of Philosophy* 8: 190–209.

Zahavi, D. (2006). "Two Takes on a One-Level Account of Consciousness."*Psyche* 12 (2): 1-9.

8. The No-Self View In Buddhist Philosophy

GORAN KARDAŠ

This me is an empirical aggregate of things objectively known. The "I" which knows them cannot itself be an aggregate, neither for psychological purposes need it be considered to be an unchanging and metaphysical entity like the Soul, or a principle like the Pure Ego, viewed as "out of time." It is a Thought, at each moment different from that of the last moment, but appropriative of the latter together with all that the latter called its own. All the experiential facts find their place in this description, unencumbered with any hypothesis save that of the existence of passing thoughts or states of mind. (William James, *The Principles of Psychology,* 1950: 400-1)

Now when the Śūnyatāvādin (propounder of emptiness or no self view, i.e. a Buddhist) attempts to communicate his doctrine he encounters this difficulty: he has a world view in which there are no essences but a language in which every item implies an essence. (Richard H. Robinson, *Early Mādhyamika in India and China,* 1967: 48)

One of the chief obstacles for modern people trying to understand Śūnyatā (emptiness) is that science discarded the substance-and-attribute mode of explanation centuries ago; and, thanks to popular science, we are all Śūnyatāvādins (propounders of emptiness / no self view) nowadays in our serious metaphysics, while often remaining naive svabhāvavādins (propounders of self-existence / own nature / self view) in our theology and self image. (Richard H. Robinson, *Buddhist Religion,* 1970: 52)

1. Language Behaviour and "Proliferated" Reality

What there is is what is "agreed upon" (*vyavahāra*) to be. There occur certain (what we call) phenomena, processes or events that can be verified by experience (understanding, feeling, volition, etc.) and communicated (language). This is Buddha's starting point and he thinks that there cannot and should not be any disagreement regarding it. Otherwise, if certain description or analysis or claim, especially the one that supposedly refers to some ultimate state of affairs "transgresses the limits of conventional us-

age" (M III: 230, 234), mental confusion inevitable arises along with meaningless (*niratthaka*) and baseless (*amūlaka*) statements (MA III: 273) and alike (philosophical) disputes. Moreover, for Buddha such statements that transgress conventional usage are utterly void of any meaning also because its adherents are not able to attach any verifiable content to it.

There are many examples in the early Buddhist canon (*Tripiṭaka* or "Three Baskets") where Buddha analyses ungrounded and baseless, mostly metaphysical or religious claims because they cannot be verified by those who propound them. For example, brahmins (priestly class) believe that they know the path which makes salvation possible and leads the one who acts according to it to a state of companionship with the highest God, Brahmā. (D. I.: 241) This claim is considered by the Buddha to be meaningless because brahmins cannot attach any meaning or verification to the term "(God) Brahmā "because they are lacking "a direct vision of Brahmā." (*Brahmā sakkhidiṭṭho*, loc.cit.) Brahmins also "do not claim to know where, whence and whither Brahmā is," but nevertheless "they claim to teach a path to the companionship of him whom they have not seen (*passanti*) or known (*jānanti*)" (ibid).

Another example of a meaningless claim because of the lack of verification is even more appealing for our discussion. Buddha analyses a statement made by brahmins that "the soul (*attā*, sanskrit: *ātman*) is extremely happy (*ekantasukhī*) and without defect (or: "healthy", *aroga*) after the ultimate death."[1] (D. I.: 192) He finds it again meaningless and baseless because those who make it cannot attach any meaning to any of the terms that constitute the statement. They have never experienced the feeling of "extreme happiness" (*ekantasukhī*) and hence are not able to attach any meaning to this expression. The same is with the "ultimate death." They did not experience it nor they could receive any reliable information from anybody about the nature of that kind of existence. (ibid.) Hence "the ultimate death" is a meaningless and contentless expression. (ibid.)

Furthermore, certain early Indian philosophers propounded, similarly to Parmenides, an ontological concept of Pure Being (*sat*, *ātman*, *brahman* as that alone and always exists by itself while all other things are merely imperfect reflections of it and consequently have only "borrowed" and temporal existence. Now, the very qualification of the Being ("always exists") directly violates the convention of time or, to be more precise, the convention of "three times" (past, present and future).[2] "Always" can only

---

[1] I.e. after the stopage the cycle of rebirth (*saṃsāra*).
[2] Cf. S III: 70-73: "There are these three [i.e. past, present and future] linguistic conventions (*nirutti-pathā*) or usages of words (*adhivacana-pathā*) or terms (*paññati-pathā*) which are distinct, have been distinct in the past, are distinct at present and will be

mean "in all times." But predicate "exists" refers to present time only, so the claim "always exists" can mean either that the future and the past exist at the present moment or that future and the past exist as the present moment. Both options are obviously nonsensical precisely because the above claim about Being is posited by the "transgressing the limits of conventional usage," in this case by transgressing the convention of a necessary (three)-time reference.

But the "transgressing the limits of the conventional usage" that provokes baseless and meaningless (metaphysical or religious) claiming and reasoning is not the only type of "unskilful" (*akusala*) dealing with the language Buddha has identified. Even for those who are sticking to the conventional usage there is a danger of "being led astray by it." There is a famous claim made by the Buddha that the "Enlightened One (lit. 'the one whose mind is free', *vimuttacitta*) is said to make use of conventional terms (*loke vuttam tena voharati*) without being led astray by them (*aparāmāsa*, lit. 'without clinging to')." (M. I.: 500) How one can be led astray by using conventional terms? Generally speaking this happens when we assume that expressions, terms and so on imply that there is corresponding objective or even ontological entity referred to by them or, to be more precise, that language items necessarily imply the objective existence of corresponding ontological entity or event.[3] The famous Buddha's example is that when he uses words like "I" (*aham*) or "self" (*ātman*) that are normal "current forms of speech" he does not imply that there exists corresponding mysterious and undying entity called "Self" that is centred in the core of all beings or so.[4] This Buddha's insight strongly reminds us of B. Russell's claim, made

---

distinct in the future and which are not ignored by the recluses and brahmins who are wise. Which three? Whatever material form there has been, which has ceased to be, which is past and has changed is called, reckoned and termed "has been" (*ahosi*), it is not reckoned as "it exists" (*atthi*) nor as "it will be" (*bhavissati*)…" (trans. Jayatilleke: 316).

[3] This does not, of course, mean that the language or language items refer to nothing, but only that the meaning and reference are constituted within the process of a language acquistion through the generations of speakers. As to technical terms, their meanings and references are constituted among "educated persons," according to ancient Indian Grammarians. Thus, according to Grammarians, e.g. the technical term "substance" (*dravya*) has a meaning that is "agreed upon" among philosophers who found it necessary to postulate such a concept. That also means that the meaning of substance, as of any other concept, is of mental nature and in no way implies its objectiveness, i.e. that there is entity named substance "out there."

[4] Buddha also gives his term *atta-paṭilābha* ("acquisition of self") as an example of an expression which may be misleading. (D. I.: 195) This term Buddha uses in the context of explanation of a certain deeper contemplative states and is of a pure phenomenological and not ontological nature. Acquistion of self actually means acquistion of a certain mental state.

some 2500 years after Buddha, that "substance in a word is a metaphysical mistake, due to the transference to the world-structure of the structure of sentences composed of a subject and predicate." (Russel 1945: 225)

Process of using expressions, designations, etc. as implying corresponding "objective" ontological entities or facts, Buddha terms as (obsessive) *proliferation* (*prapañca*). However, this process does not start within the (unskilful) language-use itself but is, according to the Buddha, the outcome of uncultivated or "unguarded" process of sense-perception, as we will see soon. Thus Buddha oftenly speaks about "notions (born of) mental proliferations" (*papañca-saññā-sankhā*, e.g. in M. I. 108)[5] that are abundantly in use in metaphysical and religious discourses.

Buddha's "conventionalism" regarding language and language-use was obviously directed against the idea that can be found in the earliest Indian cosmological speculations that there is inborn and eternal connection between language and reality or "name-and-form" (*nāma-rūpa*). In fact, according to these speculations, the nature of both is the same. Who knows a name of a thing knows at the same time a thing itself referred to.[6]

This idea persisted even within the metaphysical realism of the Nyāya and Vaiśeṣika schools (that postdate Buddha) who were all about reasoning, logic and categorial thinking. According to one influential texts from this tradition, all that can be said to exist is necessarily characterized by three things, namely that it is existent, can be cognized and can be named. This was taken very seriously. If there is a name, sanctioned by speakers as a meaningful series of phonemes, it must refer to something existent that can be cognized. Even in the case of "empty terms" such as "a son of barren woman" or "round square" they are formed on the basis of a meaningful terms (son, barren, square, etc.) but "wrongly" connected, i.e. not established by language users.[7]

Thus, for example, Indian metaphysical realists would analyze the sentence "I feel pain" and alike sentences as follows: The property (*dharma*) of pain (that inheres in the generic property "painfulness") resides in the

---

[5] Buddha terms his teaching as completely "lacking (any linguistic or mental) proliferation." (*nippapañca*, A. IV. 155)

[6] According to a myth recorded by the *Manusmṛti*, in the begining the Creator created names and states of all things from the (sacred) words of the Veda (1. 21). All things has speech (*vāk*) as their root (*mūla*), they issue from it. (4. 25); (cf. also Bronkhorst 1999: 3-7).

[7] Although Nyāya holds that the connection between words and things is not natural or inborn but is rather conventional (*saṃketika*), i.e. established through the generations of speakers, this connection once established is sanctioned by the "will of God" (NS II. 1. 55) so that "such and such meaning (*artha*) should be understood from such and such word." (*Tarkasaṃgraha*, 38)

simple substance (*dravya*) or *substratum* (*dharmin*) "Self." Here words like "I" ("self"), "pain," "property," "substance," etc. are not arbitrary words, they reflect objective (mind-independent) structure of the world. And that also apply to a well constructed sentence. It's syntax (syntactical relations) reflects objective relations (*sambandha*) that occurs among objective facts of the world. When a sentence is properly (grammatically) formed, consisting of meaningful words, there is one-to-one correspondence between language and reality. This correspondence furthermore secure the possibility of an adequate cognition of reality according to these philosophers.

But this is only a wishful thinking. What can be known and meaningfully named is only what is within the range of experience.[8] But experience is not a static feature nor can be apprehended and understand without considering non-cognitive elements such as feelings, volitions, etc.

This is the *first illusion* derived from language misbehaviour, namely that language items and syntactical relations necessarily correspond to objectively existing entities or events, discovered by the Buddha.

2. "Self" as a (Redundant) Proliferated Concept and the Impersonalization of Experience

According to Buddha, the concept of ourselves or beings in general as a principal "Self," derived from language misbehaviour, as outlined above, is the root of all our subsequent delusions and ignorance about the world. Because this concept constructs understanding that is based on something which is not there and can never be verified by experience, i.e. directly, but only indirectly, by the way of conceptualization or inferential reasoning which is not direct or immediate way of knowing.[9] For example, one might argue that the self must exist because of the evident fact of the unitary of

---

[8] Buddha oftenly stresses that the boundary of experience is at the same time the boundary of the world (*loka*). It is impossible to meaningfully talk or to think of something that is outside the range of experience. The limit of experience is the limit (at least for us) of the world: "Monks, I will teach you about 'everything' (*sabba*). And what is 'everything'? Eye and form, ear and sounds, nose and smells, … mind and mental phenomena. That is what is called 'everything'. And if, monks, somone would say: I will make known some other 'everything', that would be a groundless talk (*vācāvatthur*) from his side… And why? Because this would transgress his domain (i.e. it would be outside the reach of his experience, *avisaya*)." (S. 35. 23)

[9] Buddha clearly distinguishes his "way of knowing" from all other current in his time. Thus in M II. 211 he says that he is not a "traditionalist" (*anussavikā*, lit. "the one who is following what is heard"), nor are his insights based on mere belief alone (*kevalam saddhāmattakena*), nor is he a "reasoner" (takkī) or a "speculator"/"metaphysician" (*vimāṃsī*). He claims that he gained the highest insight (*abhiññā*) "personally" (*sāmaṃ yeva*, by himself alone, i.e. directly) into doctrine (*dhamma*) among doctrines unheard before (*pubbe ananussutesu dhammesu*).

one's experience or a memory, but these facts can be explained in different ways also, as later Buddhists have shown, without postulating the concept of self or identity.

Now, the concept of "Self," in whatever way it is constructed and defended, presupposes features like permanence, stability, unchanging basis, "identity," etc. This is because the understanding(s) of "self" is worked out within the framework of metaphysics, that is, as we have seen, according to Buddha, the outcome of language misbehaviour. For example, Indian metaphysicians (Nyāya school) tried to rationally defend the "Self" on the basis of the inherent (*samavāya*) substance-atribute (*dravya-guṇa*) relation. Properties like pain, joy, knowledge etc. are obviously attributes that must inhere in something and that can be only substance, in this case, the substance called "self." These and alike properties are, according to Nyāya, "inferential marks of Self" (*ātmano liṅgam*, NS 1.1.10.). But, the Buddha is wondering, if we somehow could remove all cognitions, emotions, perceptions, volitions, etc. from our experience, would there remain anything that is *substratum* of these properties? Most people probably believe so, but that is highly hypothetical, cannot be verified directly, and is, for the Buddha, of course, the outcome of language misbehaviour: Repeated use of personal pronoun "I" – "I do that" "I do this," "I feel," "I think" in no time will provoke obsessive proliferated concept of "Self" or "identity" to arise.

But for Buddha and the Buddhist philosophers in general, metaphysical categories are only hypostatized grammatical categories corresponding to nothing real and necessary, they are only conceptual. The Buddhist criterion for "objectively" real is very simple and in fact an empirical one: Something is objectively real if it does not depend on something else for its existence, i.e. there has to be at least some part of it that is self-established ("own/inherent nature" *svabhāva*). But what or who within the range of our experience satisfies this criterion? What or who in this world could be said to exist inherently? Obviously, nothing.

The Buddha then proposes a different kind of analysis of experience or of what we call a human being that does not presuppose any preconceived concepts or notions that would refer to some static and unchanging entities, hypostatized from language-use. He in fact completely depersonalizes experience and starts with some phenomena that can by verified by all as being immediately there. These phenomena are certain perceptions, feelings, volitions, cognizings whose causal interplay constitutes what we call experience or human being. And this kind of depersonalized (and causal) analysis should be followed by a proper type of discourse or even syntax. Here we have an example of Buddha's "linguistic turn":

Who, Venerable Sir, craves? - The question is not properly put, said the Buddha. I do not say that [someone] craves. If I had said "[someone] craves" then the question "who, Venerable Sir, craves?" would be properly put. But I do not say so. Me, not speaking thus, who would ask - "Venerable Sir, *conditioned by what craving [arises]*" – that [would be] a question properly put." "The (Buddha-) Teaching is explained/preached in a "causal" (connected) way, not in a "non-causal" (non-connected) way. (S. 2.1.2.2 and M. II. 9)

Following this new type of syntax in analysis, Buddha would analyze sentence "I feel pain" in a sense "conditioned by x, y… (feeling of) pain arises." Likewise, e.g. "I see the tree" refers to the epistemic situation where the eye sense-faculty is in the contact with the corresponding sense object ("tree") that generates corresponding (visual) consciousness (awareness that verify that said perception has occurred). And that is all there is according to Buddha. Just the occurrence of causally related phenomena. Appropriating also a later Buddhist terminology, we can say that the concept of "self" (*ātman*) is a cognitive construction (*vikapla*) or imputation (*samāropa*) formed based on the stream of psycho-physical events or "the stream of (causal) happening/becoming" (*bhavasota*).

3. How to Eliminate the Reified Concept of Self from our Analysis of Experience and Why is it of Utmost Importance?

This elimination is possible only if the main fuel that generates the concept of Self is extinguished (*nirvāṇa* = "blowing out") – that is "clinging to," "attaching to" or "appropriating" (*upādāna*) this "stream of happening/becoming."[10] This appropriation is constantly fed by a language misbehaviour that has its roots in "unskilful" (*akuśala*) cognitive behaviour, to which we are turning now.

There are many passages in the Buddhist canon where Buddha depicts the genesis of unskilful cognitive behaviour that finally ends in suffering (*dukkha*):

In dependence on visual organ and visual object there arises a (corresponding) visual consciousness. The meeting of these three is a contact. Because of the contact feeling arises. What a man feels that he sees; what he sees he reflects upon; what he reflects upon by that he is obsessed; thus illusion of "I" and "mine" arises and all that mass of suffering (*dukkha*). (M. I: 111-112) Thanks to these obsessions (*prapañca*), a man is "attacked" by obsessive perceptions and concepts regarding visual objects acquired by a visual organ, etc. (S. II.: 58)

---

[10] There are four basic types of clinging (*upādāna*) according to Buddha: clinging to sense-pleasures (*kāmupādāna*), to rituals (*sīlabbatupādāna*), to (metaphysical and/or religious) beliefs (*diṭṭhupādāna*) and to the soul/self (or substance) theories (*attāvādupādāna*). (M. I. 261) All of them has as its cause desire or "thirst" (*taṅhāpaccayā upādānam*, ibid.)

> There are, monks, signs and features of sense-objects that pass through the door of sense faculties up to mind where in the form of concepts and views generates a clinging to (*upādāna*) them and a subsequent thirst (*tāṅhā*) for them ending in the illusion of "I" and "mine" (possessor). (S. 35.: 120)

Buddha's account of the genesis of unfavourable cognitive behaviour is more or less causal. He also stresses that the cognitive is dependent upon affective. The reason why we have certain views, standpoints, believes, etc. is not because we have arrived at them by some pure rational, subjectively-detached or unbiased reasoning but quite contrary, because we are impelled to construct them as such by the force of deeply rooted (and mostly unconscious) tendencies (*saṃskārās*) that are generating from the very contact between our senses and the world in general (our social world included). For example, in some of his talks Buddha hints that the belief in Self also stems from the fear of death (i.e. fear of losing oneself). This belief is then rationalized, explained, reasoned, etc. as if completely unrelated to that fear. We could say that one's world-view is the direct product of his precognitive, immediate "sensing" of the world. The rational part ("reflecting upon") comes only afterwards but only to provoke further mental entanglements or mental "obsessions" (*prapañca*). And the vicious circle of unfavourbale cognitive behavior is closed.

Thus a "wrong view" (*mithyā-dṛṣṭi*) or "ignorance" (*avidyā*) about reality that is fueled by cognitive misbehaviour (appropriation of the stream of becoming) starts from the very contact between senses and sense-objects. We "read into" the later certain "signs and features" being attached to them, appropriating them, building after them "concepts and views," the Self being "the concept of all (nonreferring) concepts."

This is the *second illusion* that derives from cognitive (mis)behaviour discovered by Buddha.

The antidote for this cognitive misbehaviour is the cultivation of "unbiased" (appropriation-free) perceptual cognition:

> And how, my friend, a monk *guards* the door of sense-faculties? Having seen a form by the eye, he does not grasp its signs and features. But if his eye faculty is not restrained, unfavourable states of craving and discomfort might attack him. So, he practices the way of restrainment (of the eye and other sense-faculties); he guards his senses. And when his senses are thus under control, things are revealed to him *as they are* (*yathābhūtam*)., i.e. as impermanent (*anitya*) and lacking self-establishment (*anātman*). (S. 35.: 160)

This is what the Buddha calls "swimming against the stream" (*pratisrota*) of "unfavourable becoming" (*akuśalabhava*). And this "change of stream" can be practiced only through the meditative "contemplation" (*dhyāna*) or mental cultivation (*bhāvanā*) because the "swimming along the stream" is

so strongly and deeply rooted that it cannot be "redirected" by purely rational analysis or rational understanding. The basic idea here is to develop the inner capacity for unbiased, ego-detached (contemplative) analysis of various mental, emotional, volitional, etc. occurrences in terms of causes and conditions of their arising and disappearing within "the stream of happening" or "experience." This contemplative analysis[11] should finally eliminate ego-centered interpretation and understanding "letting the world to reveal itself as it is" (Buddha).

4. The World "as it is"

According to Buddha, all phenomena, whether mental or non-mental, are not self-established entities but are impermanent (*anitya*) occurrences/events that arise and disappear constantly. This constant and causal arising and disappearance of phenomena is termed by the Buddha as "dependent co-occurrence" (*pratītyasamutpāda*, lit. "co-occurrence having met"). The general idea is very simple: "This being, that becomes (*asmin sati idam hoti*)", "this not being (there), that becomes neither (*asmin asati idaṃ na hoti*)." These formulas oftenly referred to by the Buddha as *idappaccayatā* (lit. "that-conditionality") or *dhammaniyāmatā* ("the orderliness of phenomena") reveal the idea of the general conditionality that is not imposed onto phenomena (phenomena and their causal relations) but is the very nature of phenomena themselves (*dharmatā*). To analyze phenomena means nothing but to analyze specific conditions of "their" shaping or constituting and dissolving. "Phenomenon" is just a conventional name for the complex network of relations and the knowledge of these relations amounts to the knowledge of phenomena themselves. (*dhamme ñāṇam*, S. II. 58)[12]

Therefore, the Buddha is eager to warn that the concept of *pratītyasamutpāda* is not meant to be yet another (metaphysical) causal theory (produced by itself, produced by other, etc.) or causal scheme in a sense that a cause (or causes) has an ontological priority over effect or is more "real"

---

[11] Buddha is quite explicit that "normal" rational analysis cannot reach deeper structures that frame our experience. Only through unbiased meditative concentration (*samādhi*, lit. "putting thoughts together") that should be trained persistently, one can gain the "knowledge and insight" (*ñāṇadassana*) of things "as they become" (*yathābhūta*) (A. III.: 200). Conversely, when one's mind (*cetas*) is taken up by passions (*kāma*) and desires (*rāga*) unable to eliminate it, this is the cause (*nissaraṇa*) of his failure to know and see things "as they become" (*yathābhūta*), along with ill-will, sloth, torpor, doubt, etc. that accompain it. (S. V.: 127.)

[12] "'Causation' is not one thing and 'things involved in causation' another.... to be a thing is to be a causal thing, to be conditioned and a condition." (H. Cruise, "Early Buddhism: Some recent misconceptions," *Philosophy East & West* 33, 1983: 155)

than the effect or that the latter is produced by the former. And that is because both are impermanent (*anitya*) and "without support in themselves (*anātman*)." What is here at stake, according to Buddha is a "co-occurrence" or "co-ordination" (*sāmagrī*, in latter Buddhist terminology) between phenomena or events that we falsely interpret as "causal production" (cause-effect relation).

The fact that e.g. a pain occurred because of the contact with some sharp object does not require further elaboration or grounding. It just happened because of co-ordination or co-occurrence between arising of pain and the contact with some sharp object. Sharp object didn't produce the feeling of pain. Buddha here explains that "the pain is not self-caused (*sayaṁ kataṃ*) nor is caused by the other (*paraṁ kataṃ*)" (cf. S. 12. 17.); it just occurs when co-ordinated (concurrency) with certain other phenomena (that we falsely interpret as its productive "cause')."

The same interpretation applies to any event. There is no "agent" (*kartṛ*) of anything. Supposed agent (like the "Self") is only falsely "superimposed" (*samāropa*) upon "natural occurrences of phenomena" (*dharmatā*).

This is the *third illusion* (illusion of an agent underlying the processes or events) derived from ontological misbehaviour discovered by Buddha.

5. Some Further Remarks on the "Nature of Phenomena"

A Realist (at least an Indian one) would object: Impermanence and lack of self-establishment are inherent properties (*svabhāva-dharma*) of phenomena. So, the latter are nevertheless established as (ontologically) existent (*sadbhāva*). And why cannot a Self be regarded in a similar way?

Buddha and the Buddhists completely discards this kind of substance-based (property-property possessor) analysis. Feelings, perceptions, cognitions, and so on are not foundational (ultimate) albeit impermanent constituents or properties of experience, that are reached by a reductive analysis. This position was shaped within the reductionist tendencies in Buddhist philosophy after Buddha, in the so called Abhidharma Buddhist philosophy according to which there are foundational properties (*dharma*) of experience but not property-possessors (*dharmin*).

As for the Buddha's position on that matter, and it was never explicitly stated, I would say following: First of all, all of key concepts and terms he employs should be understood in a processual and not static sense. When he speaks about consciousness, perceptions, dispositions, sense faculties, and so on, what he has in mind are specific functions that are, for the sake of interpretation of experience, "abstracted" so to speak from the "flood of the becoming-stream" (*bhavasrota*), and not its ultimate bearing elements or constituents. What we have here is a constant ordered change of states

and circumstances without any fixed focal point, be it a Self or consciousness, perception, or some other impersonal element.

6. (Something like) Conclusion

1. What is impermanent and subject to change is "not fit" to be regarded as Self (a mysterious entity that centers and unifies experience). Argument: there is no any counter-argument (i.e. that anything is permanent and not subject to change) for this claim (so far)!

2. It is pointless to speak of a Self apart from experience (Buddha: "where there is no feeling at all, is it possible that one might say 'I am'?") If we can meaningfully speak about notion of Self at all, it is only regarding ever-changing "stream of happening/becoming." Therefore, it can only have "notional existence": *prajñaptir upādāya*, "a concept based on," referring to nothing actual in terms of exact correspondence.

3. The continuity of experience is explained by "dependent co-occurrence" (*pratītyasamutpāda*). Any idea of a permanent subject of experience and agent behind action, whether this is a "global" concept of individuality such as "Self" or "person," or an element or event such as consciousness, is replaced in Buddha's thinking by the idea of a congeries of impersonal, conditioned elements or events that are impermanent and lacking self-establishment. It is the combination or co-ordination of these elements or events that explains the fact of human life and experience, and its continuity.

7. Later Developments in Buddhist Philosophy (Briefly)

We find two opposed tendencies:

1. Realist-reductionist-foundationalist (abhidharma): Feelings, consciousness, etc., are real properties of experience (possessing inherent nature, *svabhāva*). Everything else is a mental construct (*vikalpa*) based on the causal interplay between these real properties. But they are impermanent, and to be impermanent (*anitya*) means to have a momentary (*kṣaṇika*) existence that is further defined as "being capable of (momentary) activity" = causal efficiency (*arthakriyāśakti*). And to be causally efficient is the only mark of the real existent. Since a supposed Self is permanent, it cannot do anything, not being reliable to change. Thus, it can exist only as a mental construct (i.e. notionally).

2. Anti-realist-anti-reductionist-anti-foundationalist (Madhyamaka, "Middle Way" school; hugely influenced East Asian Zen): Radical attack on the abhidharma concept of "inherent nature" (*svabhāva*). If you say that Self does not exist and only simple impersonal properties are existent

(reductionist position), that is nevertheless Self-position, only disguised. Whatever we can possible say *about* phenomena is "imputed" talk, especially when we try to "analyze" them, in searhing for their deeper or "ultimate foundation." We should stick to convention (*vyavahāra*) without trying to establish anything behind or beneath the scene of becoming: "things are justified/intelligible/conceivable only when not analyzed" (Jñānagarbha). The concept of Self is just a drop in the ocean of mental imputations that we constantly impose upon everything. The hidden purpose of the Buddha's teaching of "dependent co-occurrence" (*pratītyasamutpāda*) is pointing to the "emptiness" (*śūnyatā*): whatever exists in dependence upon something else, does not exist inherently, i.e., is "empty" and that emptiness is empty too (should not be reified). *Nirvāṇa* is nothing but "complete stoppage of mental proliferations imposed upon reality." (Candrakīrti)

J. L. Garfield and G. Priest gave a good characterization of the Madhyamaka "position": "Penetrating to the depths of being, we find ourselves back on the surface of things, and so discover that there is nothing, after all, beneath these deceptive surfaces. Moreover, what is deceptive about them is simply the fact that we take there to be ontological depths lurking just beneath." (Garfield, J. L. and Priest, G. "Nāgārjuna and the Limits of Thought." p. 15)

REFERENCES

A = *Aṅguttara-nikāya*, Vol. 1-5 (1885-1900). Morris, R. and Hardy, E. (eds.) PTS.

Bronhorst, J. (1999). *Language and Reality. On an Episode in Indian Thought*. Brill, Leiden.

D = *Dīgha-nikāya*, Vol. 1-3 (1889-1910). (eds.) T. W. Rhys Davids and J.E. Carpenter.

Cruise, H. "Early Buddhism: Some recent misconceptions." *Philosophy East & West* 33, 1983.

Garfield, J. L. and Priest, G. (2002). "Nāgārjuna and the Limits of Thought." *Philosophy East and West* 53 (1).

James, W. (1950), *The Principles of Psychology: In two volumes. Vol. 2.* New York: Dover Publications.

M = *Majjhima-nikāya*, Vol. 1-3 (1888-1899). (eds.) V. Trenckner and R. Chalmers.

MA = *Majjhima-nikāya Commentary* (*Papañcasūdanī*). Vol. 1-4 (1922-1937). (eds.) J. H. Woods and D. Kosambi.

Manusmṛti (2003). (ed.) and trans. M. N. Dutt, Chaukhamba Orientalia, Delhi.

NS = Nyāyasūtra.

Robinson, R. H. (1967). *Early Mādhyamika in India and China.* University of Wisconsin Press.

Robinson, R. H. (1970). *Buddhist Religion.* Wadsworth Publishing.

Russell, B. (1945). *A History of Western Philosophy.* George Allen and Unwin Ltd.

S = *Saṃyutta-nikāya*, Vol. 1-5 (1884-1904). (ed.) L. Feer, PTS.

Tarkasaṃgraha, *A Premier of Indian Logic* (1951). (ed. and trans.) Mahama-hopadhyaya Vidyavacaspati, The Kappuswami Sastri Research Institute, Mylapore, Madras.

9. The Self in Ancient Philosophy

ANA GAVRAN MILOŠ

1.

The problem of selfhood or personal identity has a long historical background, but in contemporary debate the problem is analyzed in terms of two basic questions: first, what is it to be a person or what is a self; and second, what is the criterion for a person to persist through time. The first question aims to provide conditions that have to be fulfilled in order for something to be a person, given as a list of some mental or/and bodily properties by which we can discriminate between persons and non-persons. In that sense the first questions is normative aiming to define what it means to be a person. The second questions concerns the conditions for being exactly the same person through time, that is, the conditions for the survival of persons. In other words, the so-called persistence conditions explain what it takes for the same person to continue to exist over time rather than to cease to exist.

What underlies this discussion is the presupposed intuition that human beings understand themselves as unique persons. It is *me* who has memories about *my* past, who cares about *my* future and who thinks of *my* existence. It appears that a constitutive element of such understanding of uniqueness is related with distinctively subjective and first-personal view of myself inevitable as a starting point for understanding who I am as a unique person and the way I recognize myself. Exactly this aspect of personhood characterized by the self-reflexive attitude of myself, expressed with a first-personal pronoun "I," is usually understood as *selfhood*. However, there is a general agreement among scholars that such a theoretical framework in philosophical discussion appears not before Descartes and became a standard after Descartes' formulation of the concept of mind and the first-personal understanding of the self.

Namely, central to the Cartesian picture of human personality is a self-conscious individual with a privileged access to her own mental states. So when *I* think of *myself*, it is *I* who thinks of my walking, eating, willing or thinking, from which then *I* infer a specific sense of who am I. In the center of such experience is a self-conscious subject aware of her own men-

tal states with a privileged access to her mind and understanding of herself. This introspective inquiry resulted in Descartes famous conclusion that he is *res cogitans*, that is, that his true self is identified with a concept of a thing that thinks. Descartes insists on the idea that there is a fundamental certainty in the way we are aware of our own mental processes because of which we cannot fail to know that we are thinking. Exactly this constitutes the robust idea that I am the owner of my thoughts and makes self-consciousness a basic element of understanding myself as a person. Cartesian method of self-reflective inquiry became dominant in post-Cartesian tradition and shaped this subjective or "I"–centred framework for the problem of the self, where first-personal perspective is a starting point for any inquiry on personhood. In spite of the fact that Descartes successors ended up with rather different solutions of what the self might be, many examples in the history after Descartes reflect that Cartesian assumption come to have an enormous influence and shaped the debate in modern and contemporary philosophy.

Although it has been systematically criticized already by Hume and thoroughly in contemporary analytic philosophy, the Cartesian concept of the self seems to capture our basic intuitions about understanding our selves as humans and, more importantly, as individuals. So whatever we take the self to be, the importance of "I"–perspective enters our research. Richard Sorabji (2008: 22) explains this as "a *need* to see the world in terms of me and me again" because such a perspective guides and constitutes our intentions, actions and emotions and as such is essential for our survival. More precisely, Sorabji argues that our moral agency, compassion for suffering, perceptions or development of language presupposes a thick conception of self in terms of me as an individual and owner of those experiences and those who neglect this takes the burden of proof to explain those activities without the concept of self.

It is beyond doubt that Greek philosophers were also interested in the problem of understanding our selves, our nature, abilities and place in the world. It is a widespread idea that Greeks actually postulated and shaped almost all of our philosophical problems and concepts among which is the self. However, some scholars argue that Greek philosophers did not operate with a notion of self or person in the same sense as it is described above in the post-Cartesian philosophy. Namely, Christopher Gill (1996: 2006) argues that Greek philosophers did not operate with the subjective-individualistic concept of person or self at all. It is because, in his view, Greeks never adopted first-personal point of view, but discuss the notion of self under a wider problem of what it generally takes to be a human being, focusing on objective features of human identity.

Nevertheless, such an interpretative framework of the ancient self faces a serious problem. Namely, it seems to ignore the relevance of an individual life and at least one clear subjective aspect of personhood that can be found in Greek philosophy, the ethical aspect of the debate focused on moral agents and responsibility for actions. Namely, ancient philosophers were interested in understanding of what makes one's life valuable as it is captured in a central question of Greek ethics: what does happiness (*eudaimonia*) consists of? However it seems to be problematic to neglect another aspect of the same question, namely the one that concerns an individual aspect in regards of achieving happiness: what should *I* do in order to achieve *eudaimonia*? In other words, this aspect thus puts forward an individual aspect of moral agency with more subjective interest that should be combined with a normative and objective account of happiness.

So, in what follows I will explore Gill's view of the ancient self in more details and argue that as such it does not give satisfactory answer to the problem noted above. In order to expose Gill's account I will briefly present three examples of ancient accounts of the self in Plato, Aristotle and Epicurus. Next, I will argue for a different reading of the ancient concept of the self based on Richard Sorabji's view who claims that Greeks had a various models of selfhood, among which we can also find the one similar (not identical) to the modern concept of subjective-individualist selfhood. I will also rely on Anthony Long's attempt to reconcile objective, human identity with subjectivity of an individual self, and claim that we can find both objective and subjective elements of the self in Greek tradition. Nevertheless I do not want to argue that it is possible to find traces of Cartesian understanding of the self but to put forward the idea that ancient framework of the debate offers a different understanding of personhood that does not excludes individuality, but contrary takes into account both objective and subjective aspects of the self. My aim is to show that Sorabji's arguments for such an interpretation are more compatible with the ancient ethical framework oriented toward an individual's interest in living a good life and achieving happiness.

2.

Ancient philosophers did not discuss the problem of the self or personal identity under some systematic and structured debate with clearly distinguished terms and definitions. However Greek thinkers investigated the nature of specifically human identity and the self motivated by the common general questions: Is there life after death? What is the difference between mortal human beings and immortal gods? What is the difference between living or animate and inanimate things? How can we explain sur-

vival of changes that living and non-living things undergo through time? The key notion that underlies answers to the posed questions is the concept of soul or *psuchê*.

In Greek philosophical tradition the soul is a very broad concept that generally speaking was supposed to explain the distinction between the things that are alive and capable of self-movement, and those that are not. However, there is not agreement among ancient thinkers on the more precise account of the nature of soul and the way the soul actually functions. So in some cases *psuchê* stands for a bearer of all mental states and cognitive functions common to human beings, such as reasoning, thinking, imagining, desiring, while in the other it is a specific characteristic of *all* living beings and something that very generally explains the difference between animate and inanimate things referring to the physiological processes such as growth, motion, digestion or procreation. In the metaphysical aspect the ancient debate again does not provide a unified answer about the nature of the soul. For Plato our soul is immaterial and immortal, for Hellenistic philosophers it is completely corporeal and reduced to the basic elements of reality, while for Aristotle it is a specific kind of a composition that, in the case of human beings, enables specifically human functions. Also, for Greeks the soul is something to which they ascribe moral virtues and it is the notion that explains our moral character, ability to act as morally virtuous agents within the eudaimonistic and teleological framework of ancient ethics.

Nevertheless, having all this in mind, the soul appears to be the concept that should explain, among other things, the fundamental nature of human beings, that is, the features that makes the essence of a human being and defines its identity, accounts for the persistence of a person over time and also for her unique character as one and the same person that might survive the death or maybe reappear in the future as one and the same person. The soul explains the way humans act as rational and moral subjects, capturing thus various aspects of human beings in general but also as individual subjects. Therefore, the concept of the soul in the ancient debate is the best candidate to start with in order to develop the ancient conception of self and avoid an obvious danger of anachronism. As Sorabji points out:

> The self in the ancient philosophers is seldom identical with the soul. Often it is only one aspect of soul, its reason or will, for example, or a part of soul to be distinguished from the shade or ghost. In the theories of reincarnation, the same soul may be successively borrowed by entirely different people, and so outlasts any one self. Sometimes the self is the body, or includes the body along the whole person. Although the pronouns pick out only a thin self, specifications of what the self consists in are thick, and this con-

trasts with some of the very thin conceptions of selfhood passed on us by certain 17ᵗʰ and 18ᵗʰ writers on selfhood. (Sorabji 2008b: 17)

In the quoted passage Sorabji puts forward another important idea for the paper, namely, that the ancient self is a much broader concept than the Cartesian self and more importantly, that the ancient concept of self can refer to many different things, arguing for the "astonishing variety of self" in antiquity. Among those various models, Sorabji claims, there are examples of self being identified with just one aspect of human nature, with some objective characteristics of what it takes to be a human, but also with individual self, implying the first-personal aspect of personhood. Gill, on the other side, accepts the distinction between the Cartesian and the ancient self, but however aims to argue for more objective and fixed understanding of the self in antiquity as revealing specifically normative aspect of the human nature. In what follows I will first briefly present three key examples central for Gill and Sorabji's disagreement: Plato's dualistic concept of the soul, Aristotle's hylemorphism and Epicurus' materialistic account of the soul.

2.1. Plato

For Plato the true self or the essence of the person is equated with the notion of *psuchê*. Plato accepts dualism and claims that the soul and the body are two distinct things because the soul is immaterial, immortal and able to outlive the body. He accepts the general characteristic of the soul as the animator of the body saying whenever soul takes possession of a body, it always brings life. (105c-e) The main characteristics of the soul in the *Phaedo* are given as a list of cognitive and intellectual capacities, from perceiving, desiring, feeling emotions or reasoning. However, the most important feature of the soul is that it enables reasoning emphasizing thus its rational aspect as something that regulates and controls the body, its desires and affections. (63b-c) Also, the soul is the bearer of other moral characteristics such as temperance, justice and courage and as such it constitutes a moral character of the person.

Within such a framework it seems plausible to identify the soul with the person or the self. Exactly the soul is something that explains the main aspects of the concept of self: the same person continues to exist over time since what remains the same during the change is the soul while it is embodied, but also it ensures personal survival after death and as such is one's essential self. The fact that the soul is the bearer of *personal* survival is nicely captured in Socrates' last words before he took the poison:

> "We will do our best," said Crito. "But in what way would you have us bury you?" "In any way that you like; only you must get hold of me, and take care

that I do not walk away from you." Then he turned to us, and added with a smile: "I cannot make Crito believe that I am the same Socrates who have been talking and conducting the argument; he fancies that I am the other Socrates whom he will soon see, a dead body— and he asks, how shall he bury me? And though I have spoken many words in the endeavor to show that when I have drunk the poison I shall leave you and go to the joys of the blessed—these words of mine, with which I comforted [*paramutheîsthai* = divert by way of *mūthos*] you and myself, have had, I perceive, no effect upon Crito. And therefore I want you to be surety for me now, as he was surety for me at the trial: but let the promise be of another sort; for he was my surety to the judges that I would remain, but you must be my surety to him that I shall not remain, but go away and depart; and then he will suffer less at my death, and not be grieved when he sees my body being burned or buried. (*Phaedo* 115c-e, transl. Jowett)

Socrates wants to reassure Crito saying that *he* is not his body that soon is going to be buried, but his *soul*, which is the reason why Crito should not be upset seeing Socrates' dead body. Plato thus endorses dualistic position where the soul is taken to be an essential self of a person and the bearer of personal identity. The soul in the *Phaedo* is characterized as simple (i.e. partless), immaterial, changeless and immortal and sharply detached from body.

In the *Republic* Plato is still committed to the dualistic assumption, except that now he claims that the soul is not simple, but consists of parts. An illustrative passage from the Republic is the following one:

So one who says that justice pays would claim that we must do and say what results in the man within being the strongest [part] of the man, and in his taking care, like a farmer, of the many-headed creature, nurturing what is tame and domesticating it, preventing what is wild from growing, turning the lion's nature into an ally, and nurturing by caring for all the [parts] in common and making them friendly to each other and to himself. (*Republic* 9, 589a–b)

Plato claims that each human soul consists of three parts associated with an inner human, a lion and a many-headed beast, where these correspond to the reason or intellect, the emotional and the appetitive parts of the soul. Since the reason in this metaphor is described as the man or human, it suggests that the reason then is the true man or the true self, as opposed to the irrational parts of the soul. In Plato's view the reason should dominate over the irrational parts having the best qualification for ruling over the soul: the ability to calculate what is best for the soul as a whole, and only then it is possible to achieve the best human life and *eudaimonia*. Such a picture also opens a possibility for a diversity of selves, as Long notices, within "a spectrum of self-identifications for persons, ranging from the truly philosophical right down to the fully bestial." (Long 2015: 152)

2.2. Aristotle

Aristotle rejects Plato's body-soul dualism and offers a different concept of the soul as a sum of abilities of an animated organism to perform specific functions characteristic for animated organisms. For Aristotle the soul is a distinguishing mark of all animated bodies, both plants, animals and humans, and serves as a principle that makes them living things. The living things' soul then is identified with their specific functions, that is, with the characteristic abilities only living things are able to perform. To use an Aristotelian term, the soul is the form of a specific matter that has life potentially (natural, organic body), and exactly the form is what makes this specific matter what it actually is. In that sense, particular living thing is the compound of matter (but not any kind of matter, only that has life potentially, i.e. natural body) and the form that organizes or shapes matter in such a way that actually enables it to perform specific activities typical for that particular thing.

All animated things share some basic abilities, ability to nourish themselves, to grow and to reproduce, and they are able to do that exactly because they have a soul (they are animated by the soul). The soul with only those basic abilities is found in plants and is called nutritive or vegetative soul. Unlike plants, animals and humans have additional ability to perceive, having at least the sense of touch. Animals thus have a sensitive soul that enables them to do all the things plants are capable of, plus to perceive, which also includes capacities for desiring, feeling pleasure and pain. Finally, humans have all the abilities found in plants and animals, plus an additional ability to think or understanding (*noein*) provided by intellect (*nous*), having thus a rational soul. (*De Anima* II. 2)

Since the soul is taken to be a set of capacities possessed by natural bodies, it becomes clear that the soul cannot exist apart from the body as an independent and ontologically different substance. So unlike in the Platonic account of immaterial and immortal soul, for Aristotle then the soul cannot continue to live after the death of the body, since the soul cannot exist independently of the body. According to him, a person cannot survive her death simply because she is a complex thing composed of specific matter and a typically human form. So, although the form is the principle of life, in order to be realized it has to be inseparable from the body.[1] However, although Aristotle rejects dualism it would not be correct to say that he accepts any kind of materialism. The Aristotelian theory is in fact a third theoretical framework, so called hylomorphic theory, where the soul as a

---

[1] For the explanation of the interpretation of Aristotle nous as immortal see Sorabji 1999: 9-12.

form (*morphe*) is a specific composition of matter (*hule*) that causes the unity of a single organism, i.e. a particular plant, that animal or Socrates. If we take the example of Socrates, what makes him the same person over time is the fact that he continues to exist as the same compound of his particular matter animated by the human soul and capable for rational activity. What makes him different from Democritus is primarily explained by a difference in matter, but what they have in common is distinctively human nature, that is, the rational soul. This aspect of the self we might call an ontological or biological self, explained in terms of form and matter. (cf. Sorabji 1999: 8-9)

In the *Nicomachean Ethics* Aristotle gives some further explanation of the soul relevant for the ethical discussion and practical purposes for the human behavior. Human behavior in Aristotle's view is always goal-oriented, that is, every human activity is teleological or purposive aiming to achieve an end (*telos*). (*NE* I. 1) So in order to understand what is a particularly human end we need to investigate the human soul since the soul is what enables distinctively human life and human activities. As we have already learned from *De Anima*, what makes a distinctively human soul is its rational activity or life in accordance with reason. However in *NE* Aristotle exposed his famous functional argument (*NE* I. 7) where we learned that he is not interested in any kind of human life, but the best possible one, the one that will completely realize its human purpose in achieving happiness or *eudaimonia* as the final human good. So starting from the specific human function as a rational activity of the soul, Aristotle infers that in order to live the best possible human life, we must do well what is a characteristically human function. Therefore, the rational activity of the soul has to be done in accordance with virtue or excellence.

So in the context of ethical discussion the self is explained in terms of exercising characteristically human rational activity within the practical dimension. Here he puts forward the idea of distinctively human action (*NE* III. 2) explained in terms of having a proper idea about the goal, then making a deliberative choice (*proairesis*) about the way to achieve the goal and finally deciding to act. The key notion here is *proairesis* and as Sorabji (2008: 35) points out, it serves as the best candidate for the ethical aspect of Aristotelian self since Aristotle in *NE* VI. 2 says that *proairesis* as the source of action "is the human."

2.3. Epicurus

Epicurus' philosophical inquiry was driven by our need to understand the sources of unhappiness so he recognized that the main disturbance for a happy life proceeds from false beliefs, primarily about the gods, celestial

phenomena and death. So the correct beliefs that the gods do not interfere in human affairs, that celestial bodies are not divine, as we are told in the passage, together with the belief that "death is nothing to us" will take away the fears and conduce to a good life. For the purposes of this paper I will focus on the belief about death since it provides us with the Epicurean account of the soul and immortality and to observe the self from the perspective of the fear of death.

The Epicurean concept of the soul follows the traditional understanding of the soul as a bearer of all the vital and mental functions of a living thing. Just as the rest of the Epicurean world, the soul is also made of atoms and therefore it is corporeal. (*Letter to Herodotus*, 66) It is necessary for the soul to be corporeal in order to be causally efficient and interact with the body, that is, to be able to affect bodies and be affected, rejecting thus Platonic dualistic intuition about separability of the soul. Long and Sedley explain this in terms of a rather strong body-soul interdependence since "sensation is the soul's sphere of responsibility, but it is the body that 'grants' it that responsibility, i.e. provides a suitable locus for the activity." (Long and Sedley 1987: 71) From this materialistic concept of the soul follows one important implication that the soul cannot survive bodily death and as such is not immortal. Such mortal soul in Epicurean philosophy seems to be connected with the self and personal identity. On this issue we find more textual report in Lucretius who claims the following:

> Therefore death is nothing to us, of no concern whatsoever, once it is appreciated that the mind has a mortal nature. (…) Even if the nature of our mind and the power of our spirit do have sensation after they are torn from our bodies, that is still nothing to us, who are constituted by the conjunction of body and spirit, or supposing that after our death the passage of time will bring our matter back together and reconstitute it in its present arrangement, and the light of life will be restored to us, even that eventuality would be of no concern to us, once our self-recollection was interrupted. Nor do our selves which existed in the past concern us now: we feel no anguish about them. For when you look back at the entire past span of measureless time, and then reflect how various are the motions of matter, you could easily believe that the same primary particles of which we now consist have often in the past been arranged in the same order as now. Yet our minds cannot remember it. For in between there has been an interruption of life, and all the motions have been at random, without sensation. (Lucretius, *De rerum natura*, III. 830-851, with omissions)

Lucretius makes several important points in this passage. He starts with a therapeutic goal aimed to free us from a mental disturbance caused by the fear of death, showing that fear is caused by a false belief about the nature of our soul. Once we learn the materialistic account of the mortal soul, we are on a good way to lead a happy life without disturbances. Death is noth-

ing to us because at the moment of death a subject who can experience it as bad ceases to exist.[2] One possible implication of such a view is that the self or personal identity is close to the specific atomical structure of a particular subject. This idea is further supported in the quoted passage when Lucretius considers a possible situation in which a subject's atomical configuration reappears again in the future. So, under the possibility of being back to life, should we be concerned with death?

Lucretius answers negatively for the following reasons. First, we should be concerned with the future survival only of if there would be a possibility for *us* to continue to exist as the *same* persons as we are now. However, for Lucretius this is not an option since he takes the condition for a personal survivor to be a continuity of memory. In that case any interruption of this mental continuity causes changes in personal identity. Since at death the atomical configuration of a person dissipates, it causes a break in mental continuity of that person and the same atomical configuration that might appear in the future would not be the same person. There is no possible scenario for Lucretius we should fear of once we are dead, since death is annihilation of a person.

3.

These three examples of the self in ancient philosophy, Platonic dualism, Aristotle's hylomorphic account and Epicurean materialism about the soul, give us some general insight about Greek understanding of the self as a soul. As I have previously mentioned, we can see that none of them explores the self in terms of epistemic certainty and primacy of the pure subjective self-consciousness immanent to Cartesian selfhood. What we find is a discussion of the self as the soul placed within ontological and ethical framework. The ontological self is supposed to explain the basic nature of a human being and its main function as a rational living thing. Within ontological framework, all three examples try to explain the soul in terms of what they take to be the basic constituents of reality. This ontological self varies from the immaterial soul in Platonism, to the soul as a form in Aristotle and a materialistic account of the soul in Epicureanism. All three examples also show that for ancient philosophers a proper account of a human being is related to the concept of well-being or *eudaimonia*, as a specifically human final end. This ethical self is explained in terms of the teleological-eudaimonistic framework of ancient ethics, as an ideal form of life for a rational human nature aiming to achieve *eudaimonia*. Howev-

---

[2] This argument today is known as the "no subject of harm" argument, for the first time introduced in philosophical discussion by Epicurus and his followers.

er, what seems to be puzzling is whether such a normative and objective framework allows an individual aspect of the ethical self.

My aim now is to explore an opposed interpretation in regards to the posed question offered by two leading scholars in the field, Gill and Sorabji. At the centre of their quarrel is the question whether the ancient concepts of selfhood allow any kind of subjective and individualistic element in terms of a subject's individual and egoistic interest in herself, or as Sorabji (2008b: 13) puts it, selfhood in terms of an "individual owner who sees himself or herself as *me* and *me again*." What they agree about is the fact that generally speaking there is a distinction between the ancient and the Cartesian self, since the Greek tradition obviously does not develop the concept of selfhood in terms of epistemic certainty and privileged access to one's own mental states, as it is the case with the Cartesian self.[3] However what seems to be debatable is the extent to which Greeks were interested in an individualistic and subjective understanding of the self.

Gill (1996: 2006) argues against a subjective or individualistic understanding of the self in the Greek tradition and introduces two contrasted approaches to selfhood: "subjective-individualistic" and "objective-participant." The former is a characteristic of a modern debate influenced by the Cartesian concept of the person and the Kantian moral theory where at the theoretical centre we find an individual subject "conscious of oneself as an "I," a unified locus of thought and will". (Gill 1996: 11) Within this subjective-individualistic conception, Gill claims, "to be a 'person' is to understand oneself as the possessor of a unique personal identity." (Gill 1996: 11)

On the other hand, the second conception of selfhood is established upon the notion of objectivity, which stands in the first place for the true human nature and objective ethical norms. Namely, the true or objective human nature for Gill is given in terms of the rule of reason, so the concept of person is primarily related to the knowledge of what objectively constitutes a human being. The objective aspect also takes as a key element of selfhood what is objectively taken to be the best way of human life. In that sense, the objective aspect captures the teleological-eudaimonistic framework of Greek ethics. It explains the self by appeal to the normative ideas about the essential characteristic of human nature as reason-ruled trying to achieve a final goal of typically human life, *eudaimonia*. The other, participant aspect, for the second key element of selfhood takes participation "in shared forms of human life and 'discourse' about the nature and significance of those shared forms of life." (Gill 1996: 12) This implies that social

---

[3] For a deeper analysis of the difference between ancient and modern epistemic views see Burnyeat (1980).

participation is of utmost importance for the development of human abilities and understanding oneself as a human being in contrast to the individualistic understanding of the self as an "I." Let us see now whether Gill's objective-participant framework fits with previously mentioned examples.

At first it seems that all three examples echo Gill's main assumptions since in all cases the self is primarily connected with the ontological aspect of the soul revealing thus what is an objective characteristic of human nature, i.e. reason. All three examples identify at some point the soul with the rational aspect of human nature. Next, Plato, Aristotle and Epicurus agree that such an objective paradigm of human nature should accord with the ethical ideal in terms of full realization of the potential of human nature in the practical sphere. Precondition for such realization is participation in community. As Gill points out, the participant aspect is given in Plato's educational program in the *Republic* or in the role of friendship for Aristotle and Epicurus, as a combination of normatively human perspective for which realization it is necessary to be embodied in a community with others with whom we share similar teleological aspirations.

More precisely, for example, in the *Republic* Plato claims that a necessary requirement for a moral development is an appropriate environment that enables an individual to achieve her own objective (not individualistic) good. Such an environment is established upon a specific educational system in which Plato carefully analyzes to what kind of things children should be exposed and what "should not be heard, from childhood on." (*Republic*, 386a) It follows that an individual cannot realize itself without a community that in the ideal case, as Plato sees it, consists of rational and virtuous people. For Gill this means that only in the community an individual is capable to become a person, from which he infers the essence of Greek understanding of person as:

> … the kind of animal whose psycho-ethical life (typical conceived as interplay or "dialogue" between parts of the psyche) is, in principle, capable of being shaped so as to become fully "reason-ruled" by (a) the action-guiding discourse of interpersonal and communal engagement and (b) reflective debate about the proper goals of a human life. (Gill 1996: 288)

So in order to become fully a person, that is, to realize our objective human nature, its rational capacity and ethical dispositions, we need to be a part of a community. To put it simply, the self is understood as an ideal character whose dispositions and understanding of its nature is gained through communal participation. For Gill this is an inter-personal aspect of the ancient self that is in sharp contrast with the individualistic and subjective Cartesian self.

This part of Gill's argumentation generally is not problematic. The ancient teleological-eudaimonistic framework takes social interaction as an important aspect of self-realization. Also, Gill is right in warning us to be careful in avoiding an anachronistic reading of the ancient text and shaping it in accordance with the modern concept of selfhood not possessed by the Greek philosophers. Namely, it is not questionable that insisting on the self-reflexive awareness of one's own epistemic condition as an essential characteristic of the self is not immanent to the ancient philosophers, so any interpretation of the ancient self that relies on such a notion of subjectivity is a form of anachronism. However, an overall implication of Gill's objective-participant framework is that it rejects *any* kind of individualistic or subjectivist account of the self in ancient philosophy. This seems to be rather problematic even for the examples of Plato's and Aristotle's concept of the self, which seemingly fit Gill's framework, but particularly for the Epicurean self. So, the question that remains to be answered is weather the teleological-eudaimonistic allows any subjectivity and individualism.

4.

Here I want to claim that the objective human self does not exclude an individual aspect of the self in ancient philosophy. Again, the framework is the teleological-eudaimonistic ethics, as for Gill, but my aim is to show that such a framework necessarily includes the individual aspect of selfhood. Namely, central to ancient ethics and its practical concerns is the pursuit of *one's own* happiness and the way an individual can make the best of *her own* life. So it seems that we are facing the following problem:

> There is a problem, then, in understanding the relation between subjectivity and objectivity. The problem is to explain how an individual self can see the world from a perspective which is genuinely his own but not just that of himself. Can we be both subjective and objective selves? (Long 1992: 261)

Gill answers negatively, but I want to show that the more plausible answer is a positive one. Namely, my point is that the teological-eudaimonistic framework of the self necessarily involves both objective and subjective aspects in the following way. The first one deals with our understanding of the ontology of things and learning objective things about the human nature and its fullest realization in terms of achieving what is best for a human being. The second one is more practical and concerns an individual trying to figure out the implement that insight within her own life and organize it in such a way to fulfill *eudaimonia*. In that sense, as Sorabji emphasizes, we should not choose between Gill's two perspectives, but endorse both of them since the subjective-individual one presupposes the objective-participant. (Sorabji 2008: 16) And it is possible to claim that in some examples

the objective-participant aspect is more discussed, as in the case of Plato or Aristotle, but not that the other one is completely missing, as Gill argues.

If we go back to the examples of Plato from *Phaedo* it is clear that Socrates is talking about himself in the exactly individualistic sense and not about human nature and its objective immortal characteristic, otherwise his comment to Crito would not make sense. (cf. Sorabji 2008b: 18; 2008a: 139-140) Recall that Socrates is saying that they can burry him whatever they like "provided you can catch *me* and *I* do not escape you" and trying to convince his friends not to despair since *he*, Socrates, will not die and not an ordinary human being. In the *Republic* the objective-participant aspect is pronounced at most established upon the metaphysical and epistemological realm of the Forms and clearly emphasizing the importance of the communal role for the realization of human nature and a complete subordination of uniquely individual characteristics of life. However, subordination to the objective-participant aspect still leaves a room for an individual aspect of the self.

First, as Sorabji (2008a: 117) points out, "the introduction of emotional parts of the soul makes the soul more individualistic." Next, Engberg-Pedersen in his discussion of the Stoic concept of personhood says something that might be useful for our understanding Plato:

> In exercising his [of any man] rational capacity and thus acquiring a belief about the good for man, he also becomes aware of his own rationality and comes to see rationality as a constituent of his own self. The result is that the belief about the good that he has acquired becomes a belief about his own good, and hence it becomes (at least potentially) action-guiding. (Engberg-Pedersen 2001: 123)

Exactly this description seems to be a framework for understanding Plato's philosopher-kings as described in the Cave allegory implying thus some level of individual self-realization and deliberation as the outcome of the process in which an individual philosopher-king becomes aware of his goals as a human being but then shaped his own life in accordance with such a goal, making it the goal of his own life. And only if the belief about what is normatively good becomes a belief about his or her own good, can it be action-guiding and might explain, for example, the reason why the philosopher-king leaves the best life of contemplation and returns to the cave. As Irwin (1977: 242-243) puts it, "if he is a virtuous man, he should regard public service in other people's interests as a part of the life that realizes his own happiness" and to show in practice that he cares for justice itself. (cf. *Republic* 520a-c) What we get in the end is the self constituted from all four aspects recognized by Gill: objective (rational human nature striving to achieve human final end), participant (communal participa-

tion), subjective (recognizing my own rational nature) and individual (recognizing that human end is my own final end).

The importance of the action-guiding element for understanding the individuality of the self appears again in Aristotle's notion of ethical selfhood. But before we turn to that, notice that Sorabji (2008a: 137-138) finds individuality even in Aristotle's ontological self, since every human has the same form, but is individuated by unique matter, specific to Socrates himself for example. Going back to the ethical self, let me remind you that Aristotle here explains the notion of distinctively human action (*NE* III. 2) in terms of having a proper idea about the goal (after he has established the human goal in the function argument), then making a deliberative choice (*proairesis*) about the way to achieve the goal and finally deciding to act in that particular way in order to achieve the goal. Since Aristotle in *NE* VI. 2 says that *proairesis* as the source of action "is the human," seemingly Aristotle's self is leaning towards Gill's framework.

However, Pakaluk notices that *proairesis* gives us the most information about the agent's character and, more importantly, that actions resulted from *proairesis* can be taken as a signs of character. (Pakaluk 2005: 118-151) Namely, the agent's character, i.e. her own desires, beliefs, and choices are causally related with the agent's behavior in a particular situation which serves as an expression of different individual human lives aspiring to the same human end. The central question of Aristotle's ethics is "how should I live?" and it presupposes an individual who should think of her own life plans and future interest. As it is well known, Aristotle at this point insists on thinking about our own life as a whole and to see whether one's life as a whole is directed towards the objective human end, *eudaimonia*. (*NE* I. 2) This claim makes sense only if we include an individual and subjective perspective of each person's life-plans in combination with objective characteristic of the end they tend to achieve. In order to be able to get a picture of my life as a whole, it seems necessary to have some concept of myself that continuously exist, that is, to adopt an individual-self viewpoint.

Finally, we can turn to the Epicurean examples of the self, which in my opinion seems to be the most problematic for Gill's framework. In regards to the problem of the self as we have seen, Lucretius gives us two important insights: first, that the criterion for personal identity is mental continuity; and second, that in order for a person to have any interest in her future survival, she needs to understand *herself* as one and the same person, that is, that what matters in survival is personal identity.

Gill however reads Lucretius' passage relying on Warren's (2001: 499-508) interpretation according to which "Lucretius accepts the idea that the past or future person is 'us', by the only criterion of identity offered here,

namely, the specific combination of atoms." (Gill 2006: 70) Together with Warren, Gill claims that memory *is not* the criterion of personal identity *per se*, but only the atomical configuration of a person, from which he concludes that this means again that the notion of the self is established upon a naturalistic and objective view of human nature. However, both Warren and Gill seems to be ignoring one important element in Lucretius' passage. Namely, what Lucretius aims to prove, and what Warren actually admits to be the demonstrandum of his argument is "that whatever happens after my death (and has happened before my birth) does not matter to me." (Warren 2001: 503) Again, the reason why it's not matter is *because* "our self-recollection was interrupted." The memory or the mental continuity is something that preserves our idea of me and future-me and as such is the essence of individual selfhood.

5.

To conclude, the self in ancient philosophy is different from the modern, Cartesian concept of selfhood, since it does not presuppose strong epistemic certainty as its essence. In contrast to the epistemological framework of the Cartesian self, the Greek concept of selfhood has to be placed within ontological and ethical frameworks. The ontological one gives us understanding of our nature in terms of basic constituents of the world and I focused on three representative examples: dualism, hylomorphism and atomism. On the other hand this ethical aspect explains ourselves within teleological and eudaimonistic assumptions upon which Greek ethics is established. In the paper I argue against Gill who claims that ancient selfhood is objective-participant, in contrast to the modern, subjective-individual, putting forward as its essence only objective, ideal human nature, together with a normative ethical framework and communal participation. However, the three examples I discussed showed that Sorabji's reading is more preferable since he does not replace the objective-participant with the subjective-individual, but keeps both since the second presupposes the first.

Therefore we can conclude that there is a room for individuality in ancient selfhood. The most promising understanding as I see it is a combination of individuality with the appeal to the ideal of human nature, as its objective aspect and as something we should strive for. However, I argue that in order for a person to strive for what it is objectively best for her, it necessarily includes understanding that it is the best for *me*. Or as Long excellently puts it: "We have an objective self, but we are highly selective, or should I say subjective, in how we exercise it." (Long 1992: 278)

REFERENCES

Burnyeat, M. (1982). "Idealism and Greek philosophy: What Descartes saw and Berkeley missed." *Philosophical Review* 91 (1): 3-40.

Engberg-Pedersen, T. (2001). "Stoic Philosophy and the Concept of Person." In Gill, C. (ed.) *The Person and the Human Mind: Issues in Ancient and Modern Philosophy*. Oxford University Press: 109-135.

Gill, C. (1996). *Personality in Greek Epic, Tragedy, And Philosophy: The Self in Dialogue*. Oxford University Press.

Gill, C. (2006). *The Structured Self in Hellenistic and Roman Thought*. Oxford University Press.

Irwin, T. (1977). *Plato's Moral Theory: The Early and Middle Dialogues*. Oxford University Press.

Long A. A. (1992). "Finding Oneself in Greek Philosophy." *Tijdschrift voor Filosofie* 54 (2): 255-279.

Long, A. A. (2015). *Greek Models of Mind and Self*. Harvard University Press.

Pakaluk, M. (2005). *Aristotle's Nicomachean Ethics: An Introductoin*. Cambridge University Press.

Sedley, D. and Long A. (1987). *The Hellenistic Philosophers (Vol. 1): Translations of the principal sources with philosophical commentary*. Cambridge University Press.

Sorabji, R. (1999). "Soul and Self in Ancient Philosophy." In Crabbe, M. (ed.) *From Soul to Self*. Routledge: 8-33.

Sorabji, R. (2008a). *Self: Ancient and Modern Insights about Individuality, Life and Death*. The University of Chicago Press.

Sorabji, R. (2008b). "Greco-Roman Varieties of Self." In Remes, P. and Sihvola, J. (eds.) *Ancient Philosophy of the Self*. Springer Netherlands: 13-35.

Warren, J. (2001). "Lucretian Palingenesis Recycled." *The Classical Quarterly* 51 (2): 499-508.

Ancient sources:
Plato, *Phaedo*
Plato, *Republic*
Aristotle, *De Anima*
Aristotle, *Nicomachean Ethics*
Epicurus, *Letter to Herodotus*
Lucretius, *De rerum natura*

Part IV

SELF AS AGENT

10. Ideal Self In Non-Ideal Circumstances

MATEJ SUŠNIK

1. Introduction

Any plausible theory of reasons should say something about the role that reasons play in the explanation and justification of our actions. Practical reasons, in other words, are not only expected to exert a "motivational pull" on the agent, but they should be able to justify what the agent does as well. It might seem that the explanatory challenge could be simply met by arguing that that the existence of reasons is dependent on the existence of motivation. For if motivation explains action, and if there are no reasons unless motivation is present, then the relation between reasons and explanation ceases to be mysterious. But this could hardly be the whole story. Although such a move could elegantly account for the explanatory role of reasons, connecting reasons and motivation in this way would leave the second (justificatory) challenge unanswered because it would deprive reasons of their normative force. In light of these concerns, some philosophers argue that normative reasons do not depend on one's actual motivation, but rather on the motivation one would have if one were better epistemically placed. This view is also known as internalism about reasons, and it has been under numerous attacks ever since it was firstly developed by Bernard Williams (1981).

In this paper I will try to defend Williams's internalism from one such attack. Internalists are often interpreted as claiming that one's reasons are not dependent on the motivation of one's actual self, but rather on the motivation of one's better or ideal self. But as some philosophers point out (Johnson 1999; Sobel 2001), this view overlooks the possibility that the agent's reasons may be wholly determined by the fact that the agent is not ideally placed. There could be cases in which a person can have a reason to do something although his better self would not be motivated to do that thing if he were in his place. As a result of their attempt to avoid this difficulty, internalists revise their view, but then fail to account for the explanatory role of normative reasons. My aim is to consider this objection in more detail and try to see how internalists may respond.

2. Williams's Internalism

People engage in practical reasoning because they want to determine what to do; and once they determine their reasons for action, people often act on those reasons. But common as it may be, this apparently simple process raises some difficult and unresolved questions. While it is evident that practical reasoning often results in the agent acquiring the relevant motivation, one may wonder how this could ever be the case. The problem is well known, and its roots can be traced back to David Hume's picture of reason as motivationally inert. Hume famously argued that reason is not capable of generating motivation on its own, and that some additional help is needed for that to happen. Bernard Williams (1981) expands Hume's picture, and argues that the process of reasoning can generate the appropriate motivation only if there is some motivation already present; the whole process has to start from something that is capable of moving us to action, and since only desires are suitable for this job, it is postulated that they have to exist in the background. The upshot of Williams' view is that the truth of all reason claims depends on the agent acquiring the relevant motivation as a result of rational deliberation. Simply put, according to Williams, an agent has a reason to perform some action only if he could become motivated to perform that action through the process of reasoning. And whether he could reach the relevant motivation through the process of reasoning will largely depend on the agent's present motivation (i.e. on what he actually desires). Since reasons are derived from his desires, this internalist account nicely explains how reasons get their motivating power.

Williams is well aware that it would be a mistake to make the existence of practical reasons dependent on the agent's actual desires. An agent may desire something, for example, because he holds mistaken beliefs. Suppose that I desire to drink the content of the bottle in my car because I believe it contains fresh water. But the bottle in fact contains poison, so my belief is false. (Williams 1981: 102) What this shows, Williams argues, is that I do not have a reason to drink the content of the bottle. Although I actually desire to drink it, I would lose that desire if I knew there was poison inside. Hence, making practical reasons dependent on one's desires does not yet imply that there is no place for normativity. It is precisely because reasons are dependent on one's counterfactual desires – namely, the desires one would end up having if one engaged in the process of practical reasoning – that they can be considered normative. And the explanatory role of reasons is still preserved within this picture: unless it is true that a reasoning process may lead an agent to become motivated to perform some action, it is also not true that this agent has a reason to perform that action. Obviously, as many have noticed, the force of Williams' argument largely depends on

his conception of the process of deliberation. For the purposes of this paper, however, this question is not of crucial importance, so there is no need to provide a detailed answer about what this process involves. It is enough to say that it includes "at least correcting any errors of fact and reasoning involved in the agent's view of the matter." (Williams 1995: 36) To sum up, what an agent has reason to do in his specific circumstances, Williams holds, depends on what this agent would desire to do if he did not have false beliefs and if he did not make any mistakes in reasoning.

3. Stepping Into Someone's Shoes

An agent who does not make mistakes in reasoning and whose "view of the matter" is not based on factual errors is sometimes described as someone who is "fully rational" (Smith 2004a), or as someone "who has been properly brought up" like "Aristotle's *phronimos*." (McDowell 1995: 73; Williams 1995: 189) But no matter which description is used, the key point is that this view involves the process of idealization: what a person has reason to do is determined by the motivation he would have in those circumstances if he were idealized in relevant ways. Or, somewhat differently, what one has reason to do in his specific circumstances depends on what his ideal self would be motivated to do if he were in the shoes of his actual self. This characterization opens up some important questions.

The mental exercise of stepping into the shoes of someone else is a useful tool that we often employ in ethical thinking. Among other things, an agent in this way becomes able to see the situation from a different angle, gains a better understanding of how his actions might affect others, and ultimately becomes better equipped to reach the judgment about what to do. Similarly, we might want others to step into our shoes because we are seeking advice or because we want them to feel what we feel. At least this is what we hope to achieve when we ask questions such as "What would you do if you were in my place?" or "How would you feel if you were in my place?" But the talk of imagining yourself in the shoes of someone else brings some familiar puzzles. Richard Hare nicely describes the problem: "If I imagine myself in your shoes, do I imagine myself having the same likes and dislikes as I have now, only in your circumstances; or do I imagine myself with your likes and dislikes too? Are the likes and dislikes part of the shoes or not?" (Hare 2000/1963: 126) The difficulty arises because an accurate description of one's situation sometimes cannot be given without appealing to one's personal (physical or psychological) traits. If the agent were different in some way, he would not be in that exact situation. Now, this means that in some cases one can successfully imagine oneself being in the circumstances of someone else only if one imagines having some fea-

tures of that person. But if those features are completely different from the features of a person who does the imagining, then the question is whether that implies that one really needs to imagine being someone else. (compare Hare 1981: 119) Suppose someone asks an animal torturer how he would feel if he were in the shoes of a dog whose tail is set on fire. Does that mean that this person would need to imagine being a dog? That does not make much sense. As Charles Taylor writes when considering some cases of this sort, perhaps one way to avoid these difficulties would be to provide "a theory of personal identity which could allow that two men with completely different life histories and with distinct physical and psychological characteristics might yet be the same person." (Taylor 1965: 288) But as Taylor himself observes, it is hard to imagine what such a theory would look like.

But let us go back to internalism about reasons. Is this view vulnerable to a similar objection? True enough, when we think about the relation between one's ideal and one's actual self, it is implausible to say that there are "*two* men with *completely* different life histories and with distinct physical and psychological characteristics." After all, the theory says that the truth of reason claims depends on what an agent would desire if *he himself* were idealized. Nevertheless, the problem remains. Although my ideal self is in many respects just like my actual self, it could easily happen that the shoes of my actual self do not fit him. Namely, if the essential part of my circumstances is the fact that I am not ideal, then my ideal self cannot be in my circumstances without changing them *or* without ceasing to be my *ideal* self. And if our circumstances are different, then what I have reason to do may also be different from what my ideal self has reason to do.

In order to establish this point, Michael Smith (2004a: 19) gives the example of a squash player who, due to his inability to handle defeat, forms a strong desire to hit his opponent with a racket.[1] If this man were more rational and clearheaded, he would instead desire to congratulate his opponent and shake his hand. But what needs to be taken into consideration is that he is *not* more rational and clearheaded, and that his present emotional condition prevents him from reasoning correctly and forming such a desire. Then, although this man's ideal self would be motivated to shake his opponent's hand, this is not what he, as he is now, has reason to do. Given his present circumstances, what he has reason to do is to move away from his opponent and calm down. In a somewhat different example (Johnson 2003: 574; Markovits 2011: 150; Wiland 2000: 562-3), a man whose critical thinking skills are significantly diminished due to his tendency to make logical mistakes is offered to take a logic course as a way to improve those

---

[1] The example was originally presented by Gary Watson.

skills. And again, as internalists claim, whether this person has a reason to take a logic course depends on whether his ideally rational self would be motivated to do so. But his ideally rational self is someone who by definition does not make any mistakes in reasoning, so he would not be motivated to take a logic course. Why would anyone whose critical thinking skills are ideal be motivated to improve them? This result, however, does not seem to be satisfying. For someone whose capacities for thinking clearly are impoverished, and who has a chance to become a better thinker by taking a logic course surely has a reason to do so despite the fact that his ideally rational self would lack that desire. Their circumstances are not the same, and internalism about reasons – at least in its present formulation – fails to take that fact into account.

4. The Advice Model

One possible way to deal with the above counterexamples would be to modify the internalist theory. And this is what many internalists in fact do: since they believe that the counterexamples emerge only because the theory is not carefully formulated, they argue that the best way to block them is to clearly specify the relation between one's actual and ideal self. The best-known proponent of this strategy is Michael Smith, who adopts what he calls the "advice model" of internalism. (2004a: 18–20) Since it is obvious that there are cases in which one's shoes cannot fit his ideal self, there is no requirement, as this model would state, that they should fit him in the first place. The main point is *not* to make my ideal self step into my shoes, but rather to make him look after me. My ideal self now has a job to watch my back, so to speak, and he can accomplish this by giving me advice about what to do. So what I have reason to do in my particular circumstances depends on the advice my ideal self would give me. The proponents of this model are no longer interested in what one's ideal self desires to do for himself when he is in the shoes of one's actual self, but rather in what one's ideal self desires that his non-ideal self do in his own shoes.[2] According to this revised version of internalism, therefore, my ideal self is not someone who "sets an example" I should follow, but he is rather my advisor.[3]

---

[2] Also compare Peter Railton (2003: 11). It should be noted, however, that it is not irrelevant, according to Smith, whether one's ideal self *desires* or *advises* his actual self to act in a certain way: while one's *pro tanto* reasons are grounded in the former, one's overall reasons are grounded in the latter (see Smith 2004b).

[3] The phrase "sets an example" is used in order to indicate the contrast that Smith makes with what he calls the "example model" of internalism, namely the model which is not immune to counterexamples, and which has just been discussed.

Going back to the previously mentioned cases, although it is true that the agent's ideal self would not desire to hit his opponent after the defeat in a squash game, his ideal self is perfectly aware that his actual self has this desire. And when taking into account his negative emotions and lack of self control, the agent's ideal self would certainly not advise him to approach his opponent and congratulate him. Trying to keep his actual self from falling into temptation to do something he would later regret, he would rather advise him to stay away. So this is precisely what the agent, according to Smith, has most reason to do. Similarly, it is perfectly compatible with this version of internalism to say that I may have a reason to take a logic course despite the fact that my ideal self would not be motivated to do so. The fact that he himself would not be motivated to take that course is beside the point; what is relevant is that he would advice me that *I* take that course. So once again, this is what I have most reason to do.

Since the advice model can easily cope with the counterexamples, many agree that it represents an "improvement" over the example model. (Bedke 2010: 42; Sobel 2001: 229) However, as Robert Johnson observes (1999), the advice model faces another difficulty: it fails to show how reasons and motivation are related. And that reasons are indeed related to motivation has already been pointed out by Williams: "If it is true that A has a reason to φ, then it must be possible that he should φ for that reason; and if he does act for that reason, then that reason will be the explanation of his acting." (Williams 1995: 39) It seems, however, that the advice model blocks the possibility Williams is talking about. Johnson (1999: 61–71) considers a scenario in which A wakes up one morning with a belief that he is James Bond. While we may suppose that A's ideal self would advise A to seek medical attention, the problem is, Johnson argues, that A cannot become motivated to do so. A's ideal self, on the one hand, does not hold this belief, so there is no reason for him to seek medical help, and A, on the other hand, cannot recognize that something is wrong with him, so he does not think that this reason really applies in his case. Therefore, A will remain unmoved by that reason. And even if A somehow ends up in the hospital seeking medical attention, that will not be *because* his ideal self advised him to go there. There could be a number of different reasons for which A may go to the hospital, but none of these reasons will be connected to the desires of his ideal self. The reason A needs to go the hospital is grounded in the fact that his ideal self would advise him to go there, and Johnson's essential point is that A is unable to go there for that specific reason.

5. Advice From a Different Creature

The challenge is, therefore, to provide a version of internalism that can both deal with the counterexamples and preserve the connection between reasons and motivation. The puzzle arises because it is not entirely clear how to understand the link between one's actual and ideal self. Since the circumstances someone faces could be largely determined by one's personal traits, it is no surprise that the situation I am in could be completely different than the situation my non-ideal self is in. In those cases, it appears, the result of the idealization process becomes unimportant for the actual person. For whatever my ideal self might desire, it does not need to have any effect on my particular situation. As David Sobel remarks: "[t]he idealization process […] turns us into such different creatures that it would be surprising if the well being of one's informed self and one's ordinary self consisted in the same things." (Sobel 2001: 228)

Interestingly enough, Williams anticipated this difficulty in his discussion with John McDowell (1995). Williams does not picture one's improved self as fully rational, fully knowledgeable or ideal in any other way. On the contrary, he explicitly claims that if practical reasons are analyzed in terms of the desires of some "ideal type," then reason statements can sometimes fail to be "distinctively" about some particular agent and his particular circumstances. And this is why he finds it important that such statements (such as the statement "A has a reason to φ") "say something special about A," and that they relate "more closely to the actual nature of A." (Williams 1995: 190) Reason claims should not be analyzed in terms of the desires of some *ideal* agent, but rather in terms of the desires of that *same* (actual) agent. Unless this condition is met, the counterexamples are always possible: it could easily happen that the actual agent and his ideal self have different reasons. In Williams' words, "…problems of this type can always in principle arise, until the distance between the actual and the imaginary improved agent has been reduced to zero… ." (Williams 2001: 94)

The reasons of the two selves are different because their circumstances are different, but also their circumstances are different because *they themselves* are different. The trouble is, as Sobel notes above, that the actual person and his ideal self are "different creatures." And this again raises some interesting questions. If me and my ideal self are "different creatures," then is there any relevant sense in which my ideal self is *my* ideal self? Moreover, even if he is *my* ideal self, it seems that – taking into account the circumstances I am actually in – this fact does not play any important role for me. Perhaps he is my ideal self after all, but he is just so distant from my actual self that it is irrelevant that he is *my* ideal self.

Let me briefly turn to the advice model in order to clarify what I have in mind. According to the advice model, as previously stated, what an agent has reason to do in his situation depends on what his ideal self would advise him to do in that exact situation. But when seeking advice, why should one turn to his ideal self rather than someone else who is also ideally placed? One would think that as long as the advisor is ideally placed (i.e. as long as he is in the ideal epistemic situation and makes no mistakes in reasoning), his identity is not important. Since I can receive advice from anyone who is ideally placed, my advisor could also be some neutral and impartial observer. So why is it necessary that this ideally placed advisor is *my* ideal self? What difference does it make? This remark shows that one could be doubtful whether this model falls under the rubric of internalism at all. For if my reasons do not really stem from my desires (actual or counterfactual), but rather from the advice of an ideally placed agent – an agent who could, at least in principle, be someone other than me – then in what sense are these reasons really internal?[4]

In summary, if internalists claim that one's ideal self is required to step into the shoes of his actual self, then one's ideal self cannot meet that requirement without changing the circumstances that his actual self is in, but if one's ideal self is not required to step into the shoes of his actual self and is rather pictured as his advisor, then it is no longer clear that this view deserves the internalist label. Can an internalist resolve this difficulty? I think the answer is yes, and in order to see that we need to return to Williams's version of this view.

6. *The Route Out*

Perhaps it would be useful to start this last section by giving the exact formulation of internalism, as understood by Williams. In his last paper dedicated to this topic, he formulates this view in the following way: "*A* has a reason to *φ* only if there is *a sound deliberative route* from *A*'s subjective motivational set (which I label "*S*," as in the original article) to *A*'s *φ*-ing. (Williams 2001: 91) Notice that Williams is only saying that such a deliberative route needs to *exist* for an agent to have a reason, and leaves it open whether an agent could in fact take that route from where he currently is. For, as he also notes, "[p]erhaps some unconscious obstacle, for instance,

---

[4] Johnson makes a similar point, but on different grounds. He is not concerned with the identity of one's ideal self, but rather with the fact that reasons on the advice model cannot find their path to motivation. He writes: "[I]t is misleading to present [the advice model] as a model of the internalism requirement. The latter connects reasons to motivation, but the advice model does not; it connects reasons to advice." (Johnson 1997: 621)

would have to be removed before [one] could arrive at the motivation to φ" (Williams 1995: 188). These obstacles can take various forms. If the agent is, for example, deceived – as is the case with the person deceived about the content of the bottle in his car – then the obstacle takes the form of a false belief. And this false belief, then, needs to be removed before the agent takes the deliberative route.[5]

Now it remains to be seen how Williams's version of internalism accounts for the above mentioned difficulties. Let us go back once again to Smith's example of a squash player. According to Smith's understanding of this case, as we may recall, although it is true that the agent's ideal self would be motivated to congratulate his opponent, this is not what the actual agent, due to his highly intense emotional state, has reason to do. The actual agent has reason to leave the place of the squash match. But this is not the correct description of what is going on in this case. The case of a squash player only shows that the agent presently cannot *act* on that reason, not that he does not *have* that reason. Under the assumption that there is a deliberative route from his *S* to his congratulating his opponent, then congratulating his opponent is what he has reason to do. The fact that his intense emotions present an obstacle for him to do so does not imply that his reason is removed, but rather that the obstacle itself should be removed. Perhaps it could be objected that if the agent is *not capable* of acting on some reason, then there is a sense in which he does not have that reason at all. But I do not think that it is even true to say that he is not capable of acting on that reason. It is not as if he does not have a capacity to deliberate; it is rather that he fails to exercise that capacity. And his failure to exercise that capacity is the obstacle that needs to be removed. The obstacle, in other words, comes in the form of him failing to deliberate.

What about the example of a person whose bad deliberative skills provide him with a reason to take a logic course? That example is somewhat different, but it also does not refute Williams's version of internalism. First, it could be undermined simply by attacking its basic premise – namely, that one's ideal self would not be motivated to improve his competence in reasoning because there would be no reason for him to do so. As it has already been pointed out, Williams does not think that one's reasons depend on the desires of one's *ideal* self, but rather on the existence of a sound deliberative route leading from one's *S* to him being motivated to act. If one does not deliberate ideally, then there is always room for improvement. Second, there is no inconsistency in saying that an agent could become motivated

---

[5] This interpretation of Williams is also favoured by Steven Arkonovich. He writes: "It is the possibility of deliberation *given* correct beliefs that grounds the agent's reasons, according to Williams." (Arkonovich 2011: 411, n.9)

through correct deliberation to improve his deliberation skills. For if one deliberates correctly in the practical domain, that does not mean that he also deliberates correctly in the theoretical domain. Perhaps it is even true that once the agent takes the deliberative route, he cannot reach the end of that route without being changed or improved in some relevant respects. But the internalist can give an account of that change or improvement. He may, in Williams's words, "impose … some constraints on what counts as 'deliberating correctly'." (Williams 1995: 188)

The "James Bond" example threatens to undermine any version of internalism, not only the advice model. Its basic message is that there could be reasons with no explanatory role – reasons that could motivate neither one's actual nor ideal self. An agent who suddenly forms the belief that he is James Bond has a reason to seek medical attention, but there is no way for him to get moved by that reason. For he can be moved by that reason only if he drops that belief, but as long as he holds that belief, that reason will not be capable of motivating him. Once again, it seems to me that the advocate of Williams's internalism can find the way out of this difficulty. According to my understanding of this position, whether there exists a deliberative route leading from the agent's S to him being motivated accordingly depends on whether it is possible for the agent to actually take that route. If there is an obstacle blocking the route, for example, and the agent cannot possibly remove it, then the route does not actually exist for him. Let us then assume that the obstacle blocking the route cannot possibly be removed by the agent. More specifically, assume that it is not possible for the agent to realize that his "James Bond" belief is false. In that case, it seems to me, we should say that the agent has *no* reason to visit a psychiatrist. Since he cannot possibly remove the obstacle, this means that there is no deliberative route leading from the agent's *present S* to him being motivated to visit a psychiatrist. Simply put, if he sincerely believes he is James Bond, and if he has no way of finding out that something is wrong with him, then *he*, as he is now, indeed has no reason to visit a psychiatrist.

But why suppose that it is not possible for him to start doubting that he is James Bond? What if the agent falsely believes that he is James Bond, but he also holds many other true beliefs from his past life? In that case it could be possible for him to remove the obstacle. Perhaps once he is told that he is *not* James Bond, he might be capable of realizing, through deliberation, that all his beliefs do not cohere well with the belief that he *is* James Bond, so he might in the end understand that he suffers from delusional disorder.[6]

---

[6] Although he does not directly defend Williams's model, this is in effect the proposal developed by Mark Van Roojen (2000). Van Roojen suggests that the problem disappears if one's better self is pictured as less than ideal.

So, if it is possible for him to start doubting the truth of his "James Bond" belief, then there might exist a deliberative route leading from his present *S* to him being motivated to visit a psychiatrist. In that case, the internalist should say, the agent indeed would have a reason to visit a psychiatrist. The advocate of William's internalism, therefore, should simply argue that whether this agent has a reason to seek medical attention will exclusively depend on the possibility of him realizing that he is in a delusional state.

Williams's internalism, then, not only that it has resources to cope with different counterexamples, but it also brings on the surface the importance of the process of idealization. Just as we learn something about other people when we imagine ourselves in their shoes, we also learn something about ourselves when we engage in the process of idealization. We gain a better understanding of what we really desire, what we plan to do, and what is the best way for us to proceed in given circumstances. Likewise, just as there is no point in imagining oneself in the shoes of someone else if that process implies that the agent needs to *become* someone else, there is equally no point in idealizing if somewhere along that process the agent completely ceases to be himself.

REFERENCES

Arkonovich, S. (2011). "Advisors and Deliberation." *Journal of Ethics* 15 (4): 405–424.

Bedke, M. S. (2010). "Rationalist Restrictions and External Reasons." *Philosophical Studies* 151: 39–57.

Hare, R. M. 2000. (1963). *Freedom and Reason.* Oxford University Press.

Hare, R. M. (1981). *Moral Thinking: Its Levels, Method and Point.* Oxford University Press.

Johnson, R. N. (1997). "Reasons and Advice for the Practically Rational." *Philosophy and Phenomenological Research* LVII/3: 619–625.

Johnson, R. N. (1999). "Internal Reasons and the Conditional Fallacy." *The Philosophical Quarterly* 49: 53–71.

Johnson, R. N. (2003). "Internal Reasons: Reply to Brady, Van Roojen and Gert." *The Philosophical Quarterly* 53: 573–580.

Markovits, J. (2011). "Internal Reasons and the Motivating Intuition." In Brady M. (ed.) *New Waves in Metaethics.* Basingstoke: Palgrave Macmillan: 141–165.

McDowell, J. (1995). "Might There Be External Reasons?" In Altham J. E. J. and Harrison R. (eds.) *World, Mind and Ethics: Essays on the ethical philosophy of Bernard Williams.* Cambridge University Press: 68–85.

Railton, P. (2003). "Moral Realism." In Railton P. *Facts, Values, and Norms: Essays toward a Morality of Consequence*. Cambridge University Press: 3–42.

Smith, M. (2004a). "Internal Reasons." In Smith M. *Ethics and the A Priori*. Cambridge University Press: 17–42.

Smith, M. (2004)b. "The Incoherence Argument: Reply to Shafer-Landau." In M. Smith *Ethics and the A Priori*. Cambridge University Press: 43–55.

Sobel, D. (2001). "Explanation, Internalism, and Reasons for Action." *Social Philosophy and Policy* 18 (02): 218–235.

Taylor, C. C. W. (1965). "Freedom and Reason by R. M. Hare." *Mind* 74 (294): 280–298.

Van Roojen, M. (2000). "Motivational Internalism: A Somewhat Less Idealized Account." *The Philosophical Quarterly* 50: 233-241.

Wiland, E. (2000). "Good Advice and Rational Action." *Philosophy and Phenomenological Research* 60 (3): 561–569.

Williams, B. (1981). "Internal and External Reasons." In Williams, *Moral Luck*. London: Cambridge University Press: 101–113.

Williams, B. (1995). "Replies." In J. E. J. Altham and R. Harrison (eds.) *World, Mind and Ethics: Essays on the ethical philosophy of Bernard Williams*. Cambridge University Press: 185–224.

Williams, B. (2001). "Postscript: Some Further Notes on Internal an External Reasons." In E. Millgram (ed.) *Varieties of Practical Reasoning*. MIT Press: 91–97.

11. The Disappearing Agent
FILIP ČEČ

The notion of libertarian freedom by definition invokes some kind of indeterminism in the process of decision making. The traditional libertarian thinks that without indeterminism we are merely puppets whose strings are being pulled by various deterministic processes over which we have no control. Therefore, a natural solution to resolve this deadlock is to invoke the alternative – indeterminism, and to claim that we now have the prerequisite for unchaining ourselves, to be free and morally responsible for what we do. The "only" thing a libertarian has to do now is to explain how one can have control over an indeterministic process. The libertarian thinker has invoked various solutions, but all have been criticized in the same manner: indeterminism does not help. It adds nothing, it only makes things even worse. It precludes the possibility of causal determination of an action, it makes its happening random and thus it is an unpredictable, uncontrollable, inexplicable and arbitrary process therefore it represents a dangerous add-on to the deliberation process. These worries have been voiced in various ways through various forms of the luck argument[1] which capture one specific aspect of what a critic thinks that goes wrong when one bases his account of free agency on indeterminism. In this paper I will address a specific luck argument that has been put forward against event causal libertarianism: the disappearing agent objection. I will show why some replies are unsatisfactory while dealing with this objection and, by criticizing the notion of settling and the conception of selfhood invoked by this objection I'll suggest that the event causal libertarian should reject the objection as it rests on an unacceptable ontology and that consequently, he should bite the bullet and admit that there is some residual arbitrariness in

---

[1] The incompatibility of indeterminism and free will has been criticized throughout the history of philosophy. Contemporary formulations of the luck argument are numerous and differ according to what they suggest is unavailable when we appeal to indeterminism. Some will claim that libertarian decisions do not ensure enough control (see, for example, Mele 1999), or are a matter of chance (Van Inwagen 2000), or are inexplicable (Haji 2001). For a detailed overview see (Clarke 2003, Mele 2006, Schlosser 2014).

torn decision making. In the first section of the paper I'll explain the difference between agent causal and event causal libertarianism, and I'll clarify the notion of torn decision making that characterizes some event causal accounts. The disappearing agent objection will be presented in the second part of the paper, while the various strategies a libertarian can adhere to will be presented in the third section. In the fourth section I'll analyze the notion of settling and whether it presupposes some kind of agent causal power for its realization. Finally, in the last section of the paper I'll offer what I think an event libertarian should commit himself to in order to be able to reply to the objection.

1.

Roughly speaking until the '80 of the last century the dominant libertarian views were the agent causal ones which attributed some kind of special causal power to the agent, who could, on the bases of it, bring about a specific decision without being determined to do so (Chisholm 1966; O'Connor 2000, 2009; Clarke 1993, 2003; Griffith 2010; Steward 2012). The core idea of the agent-casual account is rather simple and it boils down to the following: "a directly free action is caused by the agent" (Clarke 2003: 185); or "free will of the sort required for moral responsibility is accounted for by the existence of agents who as substances have the power to cause decisions without being causally determined to do so." (Pereboom 2014: 30) The notion of causation invoked by the agent-causalist is not reducible to causation among events involving the agent. Rather, the notion invokes an ontologically specific kind of selfhood, the agent-as-substance, an entity which has the capacity to cause free choices, and which is irreducible to event ontology. Therefore, according to most agent-causal libertarian theories the decision is up to the agent qua substance: a special form of selfhood capable of producing different outcomes in equal scenarios.

The traditional libertarian standpoint was revised and the debate was altered when novel libertarian accounts entered the arena. Accounts that do not rely on ontologically irreducible entities as the agent-as-substance or agent causation were introduced by various authors; some of them opted for an ontological framework based exclusively on states and events involving the agent and thus gave birth to what is now called event causal libertarianism. According to event causal libertarianists (Kane 1996; Ekstrom 2000; Balaguer 2010; Franklin 2011)[2] a free action will be a prod-

---

[2] Many other event causal authors could be added to the list, as well as some that embrace event ontology without committing themselves to the truth or falsity of (in)determinism as for example Albert Mele (2006).

uct of indeterministic, agent involving mental events or states which do not rely on any ontologically specific form of selfhood or specific forms of causation. The event causal libertarian will rely on event causal theories of action according to which "self-determination is to be solely analyzed in terms of, and reduced to, states and events involving the agent—such as his desires and beliefs—determining the action." (Franklin 2014: 413) According to the event causal libertarian when analyzing the causal relationship between various states or events involving the agent, and the selfhood of the agent, one doesn't have to invoke, as the agent causal framework does, a conception of the agent as irreducibly causally involved in the causation process.

The paradigmatic notion of libertarian event-causal decision making is exemplified in various instances of torn decision making.[3] The result of the torn decision making process will be a free action which will be a causal product of certain agent involving mental events which are, in part, indeterministic. Robert Kane is famous for postulating the notion of self-forming actions, decisions which "occur at difficult times of life when we are torn between competing visions of what we should do or become." (Kane 2007: 26)[4] Due to the inner struggle between two distinct sets of conflicting motives we feel torn, we experience uncertainty about what to do and consequently this uncertainty ensures that the outcome of the decision making process is not determined by influences of the past. At the same time the conflicting sets of motives will guarantee that the outcome is willed, rationally and voluntary either way we choose. (Kane 2007: 26-27) Mark Balaguer, another libertarian whose event casual account includes the notion of torn decision defines them in the following way:

> [A torn decision is] a decision in which the agent (a) has reasons for two or more options and feels torn as to which set of reasons is stronger, that is, has no conscious belief as to which option is best, given her reasons; and that (b) decides without resolving the conflict—that is, the person has the experience of "just choosing." (Balaguer 2010: 71)

An important distinction between Kane's and Balaguer's conception of torn decision is that the former defines self-forming actions as being undetermined, while the latter defines them in terms of phenomenology. Balaguer argues that we know from personal experience that we make torn

---

[3] Not all event causal libertarians adhere to torn decision making. However this essay will focus on the authors that do rely on such conception notably: Kane (1996, 2007), Balaguer (2010), Franklin (2014).

[4] Kane has presented and refined his influential notion of event causal libertarianism on numerous occasions (Kane 1996, 2005, 2007, etc.). For the purposes of this paper I rely on one of his most popular recent elaborations of his view (Kane 2007).

decisions but it is an empirical question whether they are undetermined. (Balaguer 2009: 73-74)[56] There are other important differences between Balaguer's and Kane's conceptions of torn decision[7] but for the purposes of this paper it I will use the following concept of torn decisions:

a) the agent has a feeling of being torn between two or more options;
b) the outcome is not causally determined by influences of the past (it is not a deterministically produced event);
c) the decision is probabilistically caused by agent involving events;
d) the indeterministic event is part of the decision itself, it does not happen before the process of decision making;[8]
e) the options are in a motivational equipoise: if two options are open, option A and option B, then there is 50% chance that the agent will choose option A and 50% chance that he'll choose option B.
f) the act of deciding has to be analyzed on the basis of a causal theory of action.

Let me offer an example of torn decision making. Suppose that Alberto is in love with Ernest and that they have been in a relationship for quite a long period of time. They have been keeping their relationship in secret because of the homophobic society they live in. This has been a major obstacle for both of them and their daily life has been a mess due to all the compromises and secrecies they have to adhere to and endure. Alberto, being an open person, is tired of keeping their relationship in secret and he'd like to break this prison they are in and publicly affirm their love. However, whenever he suggested something like that to Ernest, Ernest discouraged him as Ernest's job, as well as his relationship with his parents depends on maintaining this pretense. Alberto has come to an impasse. He is torn on what to and is deliberating between two options. Whether he should maintain this situation they are in, and thus prevent any possible negative outcome that might succumb them, or whether he should give an ultimatum to Ernest,

---

[5] Balaguer's account can also be read in a different way, as departing from the usual event libertarian picture and adhering to a third class of libertarian accounts, to the class of non-causal libertarianism. Derk Pereboom reads it in that way (Pereboom 2014: 36-38). In this paper I'll interpret Balaguer's position as an event-casual one.

[6] According to Carl Ginet's non-causal view an act is free when it is uncaused, it has an agent as a subject and has an actish phenomenological quality for the agent. (Ginet 1990, 1996, 2007) Other noteworthy contributions to the non-causal libertarian account have been given by Hugh McCann (McCann 1998) and Stuart Goetz (Goetz 2008).

[7] For more details see Balaguer 2010: 73-75.

[8] Some authors will say that such a decision is a directly free decision. An action is directly free just in case it is free and its freedom does not derive from the freedom of any other action. (Clarke 2011: 331)

probably something like: "either we get out of the closet or I'm leaving." Alberto's process of deliberation has raised a feeling of being torn between the options, and as both sets of reasons are of equal strength there is a 50% chance that Alberto will decide to give an ultimatum, and 50% chance that he will not act in that way. The final outcome will be made by Alberto on the bases of an indeterministic decision process grounded on the reasons he has, for the reason he has. After several days of painful deliberation he finally decides and gives an ultimatum to Ernest. According to the event causal libertarian in the exactly same scenario Alberto could have decided otherwise and opted to maintain the relationship he is in as it is.

2.

Many have argued that the indeterminism involved in the process of torn decision making undermines control. One might feel uncomfortable, to put it mildly, with a decision making process during which an agent ultimately forms a decision by "just choosing" an option between two (or more) sets of competing motives. It seems that what happens is a matter of luck. The choice that is the product of torn-decision making is in its core arbitrary.

Of course, event causal libertarians are aware of this problem and therefore rely on various solutions that grant the much needed control to the agent. Kane invokes the notion of *plural voluntary control* of the agent over his options:

> Agents have plural voluntary control over a set of options (…) when they are able to bring about *whichever* of the options they will, *when* they will to do so, *for* the reasons they will to do so, *on* purpose, rather than accidentally or by mistake, *without* being coerced or compelled in doing so or willing to do so, or otherwise controlled in doing or willing to do so by any other agents or mechanisms. (…) The conditions can be summed up by saying that the agents can choose either way *at will*. In other words, the choices are "will-setting": We set our wills one way or the other in the act of deciding itself, and not before. (Kane 2007: 30)

Balaguer adopts a similar tactic: he uses the notion of appropriate non-randomness[9] and combines it a phenomenological feeling that we experience while deciding and rather bluntly concludes:

> It is Ralph [the agent] who does the just-picking (…) at the moment of choice, nothing external to Ralph's conscious reasons and thought has any

---

[9] According to Balaguer "the central requirement that a decision needs to satisfy in order to count as appropriately nonrandom is that of having been authored and controlled by the agent in question; that is, it has to have been her decision, and she has to have controlled which option was chosen." (Balaguer 2010: 66)

> causal influence over his choice (…) Ralph chooses — consciously, intentionally, and purposefully — without being casually influenced by anything external to his conscious reasons and thought. (Balaguer 2010: 97)

Both Kane and Balaguer argue that the competing sets of reasons are the ones which provide the voluntariness and purposefulness of the resulting decision, and given that the decision is brought about consciously, by someone's own will, without any outside interference then we must conclude that the outcome is something done by, and under control of the agent.

Still, many will not be impressed by these replies as they might feel that the control expressed by the agent is insufficient. One might argue that the agent himself should be able to give the final verdict upon what to do and that the process of torn decision making does not secure that. They will insist that the agent must be the source of the decision. Moreover, someone might say that the agent isn't even present during such a decision making process. To be precise this is exactly what Derk Pereboom has in mind when he offers his version of the luck objection against event causal libertarians: the disappearing agent objection. This objection has been voiced by Pereboom on numerous occasions (Pereboom 2012; 2013; 2014; 2015) a recent one being the one presented in his 2014 book *Free Will, Agency, and Meaning in life* where he presents it in the following manner:

> Consider a decision that occurs in a context in which the agent's moral motivations favor that decision, and her prudential motivations favor her refraining from making it, and the strengths of these motivations are in equipoise. On an event-causal libertarian picture, the relevant causal conditions antecedent to the decision, i.e., the occurrence of certain agent-involving events, do not settle whether the decision will occur, but only render the occurrence of the decision about 50% probable. In fact, because no occurrence of antecedent events settles whether the decision will occur, and only antecedent events are causally relevant, nothing settles whether the decision will occur. Thus it can't be that the agent or anything about the agent settles whether the decision will occur, and she therefore will lack the control required for basic desert moral responsibility for it. (Pereboom 2014: 32)

What Pereboom wants to say is obvious: if during the decision making process an indeterministic event occurs which terminates the conflict within the agent by bringing about a decision then it is this event which does the deciding and not the agent, moreover the agent isn't even present in this stage of the decision making process, she disappears due to the fact that it is the indeterministic event the one which gives the final touch, the closure and terminates the unfortunate process by forming the decision. The agent is not present, metaphorically speaking – she disappears.

Christopher Evan Franklin while scrutinizing Pereboom's argument summed it up in the following way:

(1) On event-causal libertarianism, there is nothing about the agent that settles which decision he makes.
(2) If nothing about the agent settles which decision he makes, then the decision he makes is a matter of luck.
(3) If the decision the agent makes is a matter of luck, then he is not free with respect to or morally responsible for the decision.

Therefore,

(4) An agent who merely satisfies event-causal libertarianism is neither free with respect to nor morally responsible for any of his decisions. (Franklin 2014: 414)

Pereboom wants to stress out that one cannot be a fit subject for attribution of moral responsibility unless he possesses a specific kind of control which in turn "requires the agent to settle which of the options for decision actually occurs." (Pereboom 2012: 4) Since the torn decision making process, as formulated by the event causal libertarian, cannot secure this kind of control there is nothing left than to discard this position as unsatisfactory for attribution of moral responsibility. It must be stressed out that the objection wants to meet the control requirements as postulated by event causal libertarians. It satisfies Kane's notion of plural voluntary control as well as Balaguer's conception of appropriate non-randomness and thus maintains the idea that the agent has some kind of role in the decision making process. However, he has it only until a certain point of it, to be more specific until the point in which an indeterministic event occurs, takes over, nullifies the role of the agent and determines the outcome. That's the main reason why Pereboom thinks that the agent disappears and therefore it cannot be said that the decision was his own doing.

Alberto's case can be presented in the following manner: Alberto has two sets of competing reasons; according to one set he should give an ultimatum to Ernest, according to the other set of reasons he should keep the relationship he is in as it is. Since he is in a motivational equipoise, he cannot decide. What does the deciding is an indeterministic event which happens at the moment of equipoise, and when it happens, then the decision will be done, but Alberto will not have any influence in it since he was – undecided. Therefore, Alberto hasn't been the one who settled the matter. Alberto disappeared.

3.

There are various ways in which one may try to reply to the disappearing agent objection. In my opinion, the options available for the event causal libertarian are the following ones:

1. Formulate an answer that relies on specific events or states that fulfill the functional role of the agent. Velleman (1992) and Franklin (2014) opt for this solution.

2. Appeal to phenomenology by claiming that the decision is attributable to the agent because there is a special phenomenological feeling that only the one who decides can have. Balaguer (2010) adopts this option and perhaps Kane (2007) can be interpreted as appealing to it.

3. Devise an enriched event-causal account which will ultimately explain why the agent has not lost control over the decision making process.

4. Discard the disappearing agent objection as it relies on a form of control of the decision making process available only to agent causal theories, thus making it (1) unacceptable because of the metaphysical burden it brings along, (2) incompatible with the concept of torn decision making, (3) incompatible with the concept of motivational equipoise.

5. Bite the bullet and stick to the idea that something gets lost if one adheres to event causal libertarianism.

A counterargument to the disappearing agent objection usually boils down to a combination of several options from the list given above. However, one might rely just on one option as for example Velleman does. He opts for the first solution and argues that the agent does not disappear from the decision making process as he has identified himself with an attitude with which he is functionally identical. The attitude the agent has identified with, in Velleman case, is "the additional motivating force of the desire to act in accordance with reasons." (Velleman 1992: 479) Velleman's reply can be exemplified in the following manner: Alberto has calculated the strengths of his reasons and he noticed that he favors the ultimatum scenario. Previously he has identified himself with the desire to act in accord with his reasons. The identification brings about that the set of ultimatum reasons is additionally reinforced by the desire to act in accord with his reasons. Thus he is able to break the deadlock, ends the torn decision making process and he decides to give an ultimatum to Ernest. As long as Alberto is identified with the desire to act in accordance with his reasons, that desire will be part of Alberto's decision making process, it will fulfill Alberto's functional role as the one who does the deciding and therefore

will guarantee that the outcome can be attributed to him. It might seem that the indeterministic event nullified Alberto's control over the decision making process, but in fact, that event is only a part of the whole process which functionally is Alberto's own doing.

However, the above mentioned reply is confronted with a serious flaw. As Runyan relying on Pereboom (2015), correctly, in my opinion, stresses out:

> When a person is in motivational equipoise concerning her alternatives she is on the fence about, and out of resources for settling, what to do. There is no desire, attitude or preference in favor of one alternative. (Runyan 2015: 1634)

When an agent finds himself torn between two or more options, then all the reasons are already included in the deliberation process and that is why the resulting situation is a situation of motivational equipoise. There is nothing additional that the agent can add to the equation, no desire has been left out, adding a desire or desires to act in accordance with one's own reasons is not possible as these are, if they exist, already included in the deliberation process. What is suggested, to rephrase, is that if we are in a motivational equipoise the desire to act with one's reasons is already part of the motivational system and therefore cannot function as the "thing" that puts an end to the equipoise. Nothing can tip the scale. The only thing that can resolve the issue, the situation of being torn between two or more options, is the indeterministic event. The fifth condition of the definition of what a torn decision is prevents Velleman's solution from functioning. In a different scenario in which the chances of the available options are not tied, a scenario in which the agent is not in a motivational equipoise, the desire to act in accordance with reasons will successfully bring to a decision that can be attributed to the agent. If Alberto's reasons for handling out an ultimatum to Ernest have a 60% chance of happening then adding the desire to act in accordance with reasons would make that particular set of reasons occur by raising the probabilities to a 100% chance of happening. But this scenario cannot arise in the case of torn decision making.

Franklin combines two options from the list given above: he uses and improves Velleman's account and offers an "enriched" event causal account that includes a reductive theory of self-determination in which "the activity of the self-determining agent is reduced to a state or event that plays the self-determining agent functional role" and that "in so doing *counts* as *his* playing his functional role." (Franklin 2014: 418) In this way we have a reductionist picture of the agent, one to which Velleman doesn't commit himself, in which the states or events that count as his play the role of the settler. The states or events of relevance here are similar to the ones employed by Velleman:

> [the agent] plays a causal role *over and above* the causal role played by his desires and beliefs for action, and this supplementation amounts to his "throwing his weight" behind the desires and beliefs that led to action. It is this additional participation of the agent in action that transforms mere action into self-determined action. (Franklin 2014: 423)

Is the motivational equipoise a problem for Franklin? It depends on the interpretation of what "throwing his weight" actually means. The equipoise doesn't have to be a problem if we interpret the idea of "throwing our own weight" as a will to settle the standstill even if we are undecided. Perhaps it would amount to Alberto saying: "I really have to resolve this state I'm in. I have no inclination towards one of the alternatives but I'll pick the ultimatum option and we'll see what happens." However in this case a certain degree of randomness will remain. I doubt that Franklin would adhere to this solution as it seems that the phrase "throwing his weight" must be read in a different manner. Let me explain.

Perhaps the idea of "throwing his weight" should be read in a manner akin to Velleman's interpretation? Then the settling would be done according to the desire to act in accordance with the best reasons one has. We find plenty of passages in Franklin in which he implores this idea. For example:

> On my account, in addition to the desires and beliefs for action playing a causal role, the desire to act in accordance with the strongest reasons—a desire that is functionally identical to the agent and with which he is identified—also plays, or could have played, a causal role. It is in light of this additional causal role that the agent determines, or could have determined, and thus settled, what he would do. (Franklin 2014: 427)

But then Franklin's response is inadequate because his theory, like Velleman's, seems unable to give a proper reply to Pereboom's critique for the same reasons that were present in Velleman's case. I'm inclined to think that Velleman's and Franklin's suggestions fail to fulfill the role that the disappearing agent objection seems to require: the role of settling. In the fourth and fifth section of this paper I will argue that this role cannot be fulfilled in an event causal universe, and that fulfilling that role is a futile job, a job that the event causalist shouldn't even try to adhere to. However, before doing so I'd like to explore another option that the event causalist might appeal to when replying to the disappearing agent objection.

In his 2010 book *Free will as an open scientific problem* Mark Balaguer presented his event casual theory which has been directly criticized by Derk Pereboom several times. (Pereboom 2012, 2013, 2014) Balaguer addresses the disappearing agent objection in a recent paper (Balaguer 2012) and offers a very detailed reply to Pereboom's critique. He begins his argumentation by arguing that:

(A) Ralph's [the agent's] choice was conscious, intentional, and purposeful, with an actish phenomenology (…)
(B) the choice flowed out of Ralph's conscious reasons and thought in a nondeterministically event-causal way; and
(C) nothing external to Ralph's conscious reasons and thought had any significant causal influence over how he chose. (Balaguer 2012: 10)

According to Balaguer if these conditions are present then we may conclude that the agent authored and controlled the decision, they are indeterminate but appropriately non-random (Balaguer 2010: 66). However he is aware that such a reply might not satisfy Pereboom:

> One of the central claims in Pereboom's disappearing agent objection is that authorship and control require the agent to settle the matter. I am OK with that way of putting things. But it seems to me that if the event that settles the matter is the agent's conscious decision, then, at the very least, there is a sense in which the agent does settle it. There might be other senses—most notably, agent causal senses—in which the agent does not settle it. (Balaguer 2012: 14)

And this is exactly what Pereboom has in mind when he replies to Balaguer by saying:

> The objection is not that agents will have no causal role in producing decisions, but that the causal role that is available to agents will be insufficient for the control moral responsibility demands. (Pereboom 2013: 27)

Basically the disappearing agent objection boils down to the assumption that whichever answer an event libertarian might produce, the control that he envisages will not be thick enough. Pereboom's objection is grounded on the idea that the notion of control that the event causal libertarian is offering is simply too thin due to the nature of the decision making process the event causalist is relying upon: the torn decision itself. The causal role of the agent that various event causal libertarianists try to secure will result as insufficient if one or a combination of the first three options from the list given above is chosen. Therefore, it seems obvious that the disappearing agent objection must rely on a different, stronger conception of control, a kind of control that can be secured only by agent causal theories. If this consequence hasn't been made clear by the discussion that I provided so far, it will be crystal clear once we analyze the following quote concerning the force of the disappearing agent objection when applied to Balaguer's event causal proposal:

> But the one concern is that if the just-choosing is what secures Ralph's control, and control is a causal matter, then what is being specified is that a causal relation obtains between Ralph himself and the decision. However, the event-causal libertarian allows only causal relations among events, and not a fundamental causal relation between agent and event. (Pereboom 2014: 36)

Obviously the fundamental causal relation between agent and event that Pereboom has in mind boils down to an agent causal form of decision making. Franklin formulates this worry explicitly:

It is hard to read this objection as anything else but a bald assertion that the agent qua substance must fundamentally cause his decision, and if he does not, then he does not play the role that is required of him in free action.

> Therefore, it seems that a confutation of the disappearing agent objection must rely on the fourth or the fifth option from the above given list. Enriching the event causal account doesn't seem to suffice. (Franklin 2013: 427)

4.

The fourth option is grounded on the presumption that the disappearing agent objection can be rejected as it relies on a form of control available only to agent causal theories. A form of control that is, supposedly, indispensable for the attribution of moral responsibility but unavailable to the event causal conception of the agent. Why is it so? What is it that the agent should be capable of doing in order to be an eligible subject for the attribution of moral responsibility? What kind of control does the objection presuppose? A closer look to at the notion of settling as used by Pereboom and other authors will help to understand what kind of causal powers an agent must possess in order to have that form of control, and what are the necessary metaphysical presuppositions that render this particular form of control possible. Pereboom argues the following:

> The disappearing agent objection counts against the supposition that this [event causal] account secures the control required for moral responsibility. Intuitively, this sort of control requires *the agent* to settle which of the options for decision actually occurs. (Pereboom 2012: 4)

What Pereboom wants to say when he claims that an agent settles which option will actually occur? Broadly speaking it can be said that settling the matter implies a definite resolution of a situation in a certain way by choosing between different options available to the agent. Therefore, it seems that the notion of settling the matter is inconsistent with universal determinism.[10] Helen Steward nicely illustrates this claim by means of the following argument:

> For example, if an utterly deterministic process leads via a successive chain of causes $c_1 \ldots c_n$ to effect e, then c_n cannot count as having settled that an event of any of the types that e instantiates occurs, even though c_n is es-

---

[10] This claim has been disputed, for example by Clarke (2014), however this issue has no relevance for the purposes of this paper hence I will not tackle with it.

sential to the occurrence of e, since it was already settled at the time of c_1's occurrence that e would occur. An event can only settle a matter at the time at which it occurs, if that matter is not already settled before that time. (…) If there is ever any settling of matters in time, then universal determinism cannot be true, since according to universal determinism, everything is already settled at the start (whatever exactly we are to understand by "the start"). (Steward 2012: 40)

It is obvious, from the previous passage that the notion of settling is libertarian in its core because it requires open futures. By definition, acting by settling requires choosing between options that are available to the agent to choose from. Or as Steward expresses it "an agent's action 'just is' a matter of it being the settling of at least one from a range of possible other things that are up to the agent." (Steward 2012: 36) Furthermore, the notion of settling as used by Steward and Pereboom is not only libertarian, in the sense that it requires open futures it is also a metaphysically extremely demanding one. It seems that it presupposes some kind of agent causation because:

> … one cannot hope to analyse what it is for an agent to act in terms merely of the causation of her bodily movements by various of her mental states, because her action has to be a part of this story, the part that connects those non active mental antecedents to her bodily movements. It is the agent who has to settle the question whether those mental antecedents will result in a movement or not. That is the way commonsense psychology tells the story of action, and it cannot be retold at this level of ontology without her participation. (Steward 2012: 65)

From the quote given above it is evident that the selfhood as used by event causal theories cannot possibly settle because it is built upon agent involving mental events or states which, according to Steward do not count as an eligible possibility when we try to analyze the notion of acting. Only an *agent* can. The agent does not bring about an event according to her picture of agency but rather:

> What, properly speaking, is up to me in the usual sort of action situation is not a particular event, but rather the answers to a whole range of questions that are settled by my action when I act. (Steward 2012: 37)

Therefore, the self that is compatible with the notion of settling, the self that can settle must be one of a non-reductive cast similar to the ones presupposed by agent causal theories: a mover unmoved, an agent-as-substance. Randolph Clarke's interpretation of Steward's account can be used to further explicate some important details of her account:

> (S) An action *a* that is performed at time *t* settles at *t* whether *p* iff (i) either it is impossible that *a* be performed then and the actual laws of nature hold and *p*, or it is impossible that *a* be performed then and the actual laws hold and not-*p*, and (ii) there is nothing existing at any time *t'* prior to *t* such that

either it is impossible that that thing exist at *t'* and the actual laws hold and *p*, or it is impossible that that thing exist at *t'* and the actual laws hold and not-*p*. (Clarke 2014: 522)

Clarke offers the following example of an action that will bring to the settling of a matter:

…when I raise my arm at a certain time, *t*, my action of raising my arm might settle at *t* whether I raise my arm then, provided that nothing prior to that time suffices for its being the case that I raise my arm then. And given that my action of raising my arm settles at *t* whether I raise my arm then, it might be said that I settle at *t* whether I raise my arm then, and that I settle this matter at *t* by raising my arm then. (Clarke 2014: 522)

Evidently, the action of settling cannot be reduced to event ontology. The thing that does the settling is not an event or a state of affairs that involves the agent, but rather the agent himself: a non-reductive notion of selfhood. It is obvious that the concept of torn decisions as usually employed by the event causal libertarian is incompatible with the concept of settling for the following reasons:

(1) There is no "I" who does the settling. In the reductionist ontology of the event causalist agents qua substances are inexistent. The objection requires more than the event causal picture can possibly offer.

(2) The concept of settling contradicts with the concept of torn decision making. One important characteristic of settling is that nothing prior to the settling itself suffices for the production of the action. Nothing prior to the raising of my arm *suffices* for the raising of my arm, as Clarke invites us to think in his example. However, the concept of torn decision making implies that the situation of motivational equipoise is *sufficient* for the production of a decision. Nothing more can be added to the picture. Therefore we have two concepts that do not combine: according to the concept of settling nothing prior to the act of the agent suffices for the act, while according to the concept of torn decision making the situation of motivational equipoise suffices for the act.[11]

(3) The second point can be further expanded. It seems strange to say that the situation of motivational equipoise should be resolved by the agent. The point is that the agent himself is in such a state and he does not prefer one option over another. If he had a preference then he would

---

[11] It is important to stress out that according to the event causal picture the equipoise will be resolved by the indeterministic event of decision making. The notion of torn decision making is a process that lasts in time and incorporates both the motivational equipoise and the indeterministic event of resolving the issue by deciding. There are no interventions brought about by the agent or some other event external to the agent that could lead to what the disappearing objection wants: settling.

not be in a state of motivational equipoise. Therefore a question arises: how could the agent, even an agent-as-substance end the motivational equipoise?[12] For what reasons? What would motivate him to choose option A over B? It seems that the reason for settling is inexplicable. If that is the case then the agent causal libertarian is in the same trouble as the event causal one.

The final point leads to a further problem that is usually invoked against agent causal theories: they fail to resolve the problem of luck.

Agent causal theories employ metaphysically problematic concepts in the sense that they appeal to specific, peculiar entities such as special forms of causation or the agents-as-substances that are irreducible to other entities that inhabit the world we live in. By doing so they add novel *kind* of entities to the ontological inventory of what there is. Why? The reason is simple: then we can explain the specific process of decision making and discern it from the usual casual pathways present in the world. However why should we invoke a solution that has dubious metaphysical implications if it adds nothing?[13] Why burden our ontology with peculiar kinds of entities if these entities do not help us resolve the problem we are dealing with, in this case the problem of libertarian luck? Balaguer, among others questions what is to be gained if we appeal to agent causal theories:

> I would just like to offer one quick argument against the idea that authorship and control should be thought of as requiring agent causation. Let me put the argument in the form of a challenge to advocates of agent-causal analyses of authorship and control. The challenge is to say what exactly is to be gained by requiring agent causation. On the view I have in mind, we say that Ralph [the agent] authored and controlled his decision (…) because (roughly) the event that settled which option was chosen was the conscious

---

[12] As seen in the third section of this paper Pereboom suggests that Velleman's and Franklin's suggestions fail to meet the task given by the disappearing agent objection as the agent is in motivation equipoise and therefore a desire to decide in accordance with the strongest set of reasons will not do the trick. Something else must be added so that the agent can settle the matter. But the problem might be that nothing that can settle can be added, at least not in the event casual ontological picture. One solution that might be offered is for the agent to identify himself with a desire to end the decision process if it ends in motivational equipoise then and there by deciding. According to this proposal the agent would be the one to functionally end the equipoise however a residual arbitrariness would remain as no set of reasons would be preferred over the other.

[13] Various criticisms have been invoked along that line of thought: why should we encumber our ontology with dubious notions of selfhood such as the "substance-as-a-cause" or specific forms of causation as agent causation. In his book *Libertarian Accounts of Free Will* Randolph Clarke offers an overview of various criticisms that have been voiced against the agent-causal view. (Clarke 2003: 185-212)

decision itself. If you demand that Ralph caused option O to be chosen (or that he causally settled which option was chosen), then it seems to me that you have gained nothing; you have simply moved everything back a step. For now there is a second event, on top of the conscious decision—namely, the event of Ralph agent-causing the decision—and we can ask the very same question about this event that Pereboom wants to ask about the conscious decision; that is, we can ask what caused the agent-causal event to occur. And, of course, the agent-causal response is going to be that nothing caused it to occur. (Balaguer 2012: 14)

Balaguer's argument echoes a worry that has been present in the debate for a long time. A worry that Gary Watson nicely elucidates: "Agent-causation simply labels, not illuminates, what the libertarian needs." (Watson 1982: 10)

According to the argumentation given above the agent causalist doesn't fare much better than the event causalist. On the contrary! If he cannot explain why and how an agent settles then his position is worse than the event causal one because he has committed himself to a much richer ontology. If that is the case, then we should adhere to a less demanding ontological framework: the event causal one.

I do not want to offer a knock-down argument against agent causal theories. This is not the purpose of this paper. What I want to stress out when invoking these arguments and the questions that have been put forward is twofold. On one hand, what is suggested is that the event causal libertarian can ignore the disappearing agent objection and, on the other hand it should suggest that there are no reasons why he should follow the agent causal route and/or adhere to the metaphysical standards imposed by such a theory. It doesn't help. These arguments should motivate one to explore another possible route, the event causal one, and accept its limitations.

5.

This brings us to the fifth option that an event causal libertarian can adhere to when trying to reply to the disappearing agent objection: he should bite the bullet and stick to the idea that something gets lost if one adheres to event causal libertarianism.

As exposed in the previous section of the paper adhering to the standards imposed by the concept of settling is not something that can be achieved in an event causal ontology. Therefore the answer to the question "Should an event causal libertarian settle?" is simple: no. An event causal libertarian should explain how a free decision looks like, why it is attributable to an agent, explain the functional role of the agent's in it, how it is

incorporated in other parts of his mental life, etc.[14] He has to do so in order to demonstrate that the residual arbitrariness isn't an obstacle in the production of free decisions and attribution of moral responsibility.

On the other hand he should admit if a decision making process is grounded on probabilistically caused agent involving events then there will be some residual arbitrariness present in the decision making process. Is that an obstacle? Does the agent disappear? In order to answer that the libertarian must ask himself why does he adhere to indeterminism in the first place? What can an indeterministic world offer? One obvious answer is that such a world provides open futures in which an agent can create or follow novel causal pathways. But as already seen indeterminism is a dangerous toy to play with. Following that path does not mean that the journey will be without perils. No wonder that Randolph Clarke spoke of it as a of horror story. (Clarke 2011: 331) The horror of indeterminism is even more disturbing if one adheres to event ontology. No noumenal selves, selves-as-substances or special forms of causation are available and because of that the possibility of settling is precluded in such a world. However, the world picture of the event causal libertarian isn't that frightening, after all it is a parsimonious and intelligible ontology: no special, unique, non-reductive and unexplainable, agent-as-substances entities are being invoked. Then again, there is no settling, at least not as demanding as the disappearing agent objection requires. Consequently some residual arbitrariness will be present in the decision making process because it rests on the idea of motivational equipoise which gets resolved by an agent involving indeterministic event. The indeterminism is here to stay but to implement Balaguer's term it will be appropriately non-random. (Balaguer 2010: 66) It will be the agent's own doing.[15] Alberto will not disappear.

Therefore the event causal libertarian should bite the bullet and assume a humble approach by admitting that there is a bit of residual arbitrariness in his ontology. Toying with indeterminism demands a price to be paid.

---

[14] Some of these questions, without adhering to event causal libertarianism, are addressed in Malatesti and Čeč (forthcoming)

[15] More can and should be said regarding the issue why the torn decision making process is attributable to the agent. In my opinion an argument akin to the one presented in the eleventh footnote, the identification with a desire to end the decision making process no matter what could be good candidate, however that goal Is beyond the scope of this paper.

REFERENCES

Balaguer, M. (2010). *Free Will as an Open Scientific Problem*. MIT Press.

Balaguer, M. (2012). "Replies to McKenna, Pereboom, and Kane." *Philosophical Studies* (1): 1-22.

Chisholm, R. M. (1966). *Theory of Knowledge*. Englewood Cliffs. N.J. Prentice-Hall.

Clarke, R. (1993). "Toward a Credible Agent-Causal Account of Free Will." *Noûs*. 27 (2). 191-203.

Clarke, R. (2003). *Libertarian Accounts of Free Will*. Oxford University Press.

Clarke, R. (2011). "Alternatives for Libertarians." In Kane, R. (ed.) *The Oxford Handbook of Free Will*. 2nd edition: 329-48.

Clarke, R. (2014). "Agency and Incompatibilism." *Res Philosophica* 91 (3): 519-525.

Ekstrom, L. W. (2000). *Free Will: A Philosophical Study*. Westview.

Fischer J. M., Kane R., Pereboom, D. and Vargas, M. (2007). *Four Views on Free Will*. Wiley-Blackwell.

Franklin, C. E. (2011). "Farewell to the Luck (and Mind) Argument." *Philosophical Studies* 156 (2): 199-230.

Franklin, C. E. (2014). "Event-causal libertarianism, functional reduction, and the disappearing agent argument." *Philosophical Studies* 170 (3): 413-432.

Franklin, C. E. (2016). "If Anyone Should Be an Agent-Causalist, Then Everyone Should Be an Agent-Causalist." *Mind* 125 (500): 1101-1131.

Ginet, C. (1990). *On Action*. Cambridge University Press.

Ginet, C. (1995). "Reason's Explanation of Action." In O'Connor, T. (ed.) *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. Oxford University Press.

Ginet, C. (2007). "An Action Can Be Both Uncaused and Up to the Agent." In Lumer (ed.) *Intentionality, Deliberation, and Autonomy*. Ashgate: 243-255.

Ginet, C. (2008). "In Defense of a Non-Causal Account of Reasons Explanations." *Journal of Ethics* 12 (3/4): 229-237.

Goetz, S. (2008). "My Way." *Faith and Philosophy* 25 (2): 221-226.

Griffith, M. (2010). "Why Agent-Caused Actions Are Not Lucky." *American Philosophical Quarterly* 47 (1): 43-56.

Haji, I. (2001). "Control Conundrums: Modest Libertarianism, Responsibility, and Explanation." *Pacific Philosophical Quarterly* 82 (2): 178-200.

Kane, R. (1996). *The Significance of Free Will*. Oxford University Press.

Kane, R. (2005). *A Contemporary Introduction to Free Will*. Oxford University Press.

McCann, H. J. (1998). *The Works of Agency: On Human Action, Will, and Freedom*. Cornell University Press.

Malatesti, L. & Čeč, F. (forthcoming) "Psychopathy, mental time travel and self-knowledge."

Mele, R. A. (1999). "Kane, Luck, and the Significance of Free Will." *Philosophical Explorations* 2 (2): 96-104.

Mele, R. A. (2006). *Free Will and Luck*. Oxford University Press.

Mele, R. A. (2009). *Effective Intentions: The Power of Conscious Will*. Oxford University Press.

O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. Oxford University Press.

O'Connor, T. (2009). "Agent-Causal Power." In Handfield T. (ed.) *Dispositions and Causes*. Oxford University Press.

Pereboom, D. (2012). "The Disappearing Agent Objection to Event-Causal Libertarianism." *Philosophical Studies* 1: 1-11.

Pereboom, D. (2013). "Skepticism About Free Will." In Caruso, G. (ed.) *Exploring the Illusion of Free Will and Moral Responsibility*. Lexington Books: 19-40.

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford University Press.

Pereboom, D. (2015). "The phenomenology of agency and deterministic agent causation." In Pedersen, H. & Altman M. (eds.) *Horizons of Authenticity in Phenomenology, Existentialism, and Moral Psychology*. Springer: 277-294.

Runyan, J. D. (2016). "Events, Agents, and Settling Whether and How One Intervenes." *Philosophical Studies* 173 (6): 1629-1646.

Schlosser, M. E. (2014). "The Luck Argument Against Event-Causal Libertarianism: It is Here to Stay." *Philosophical Studies* 167 (2): 375-385.

Steward, H. (2012). *A Metaphysics for Freedom*. Oxford University Press.

Van Inwagen, P. (2000). "Free Will Remains a Mystery." *Noûs* 34 (14): 1-19.

Velleman, J. D. (1992). "What Happens When Someone Acts?" *Mind* 101 (403): 461-481.

Velleman, J. D. (1996). "The Possibility of Practical Reason." *Ethics* 106 (4): 694-726.

Watson, G. (ed.) (1982). *Free Will*. Oxford University Press.

12. Agency and Reductionism about the Self
MARKO JURJAKO

1. Introduction

When thinking about the identity of the self, we are usually thinking about issues related to the problems of personal identity. Eric Olson (2002) distinguishes between several questions that relate to the problem of personal identity. For instance, one of the questions concerns the problem of identification and can be expressed as: who are we? Here an answer could be that one part of my identity is that I am politically a leftist and I identify myself with a group of people who hold liberal political views.

Other important questions related to personal identity include problems of defining the persistence conditions of a person or a self and of deciding what it is that we are or what we are identical with. The former refers to the conditions that, for example, determine when one person at time t is identical with a person at some time before t. The latter problem pertains to the issue what we are made of. For example, some people argue that we are essentially human animals or organisms. (DeGrazia 2005, Olson 1997, Snowdon 2014) Others argue that we are material beings that are constituted by an organism, but are not identical with it, rather our identity is determined by appropriate psychological connections. (Baker 2000, Johnston 1987) Some authors argue that we should be identified with parts of our bodies, such as our brain (McMahan 2002), while still others argue that we are immaterial souls (Swinburne 1984).

Besides the metaphysical aspects, problems of personal identity are important for us because of their practical implications. (Shoemaker 2007) For example, if a person X stole something from a person Y, then it is important for us to know that a person Z at time t is the same person as X at the time of the stealing, so that we can inflict on X a just punishment. Similarly, it seems especially important for anyone to know whether she will be the person who will experience pain tomorrow when visiting a dentist or whether somebody else will experience similar pain.

The contemporary discussion on personal identity seems to be focusing on two broad issues. (Bělohrad 2014a, Schechtman 2014) On the one hand, philosophers try to give an account of personal identity that can vindicate

our practical concerns. These investigations, thus, concern issues that are related to responsibility, blame, prudential concern, moral rights, and so forth. (Shoemaker 2007) On the other hand, some philosophers hold that strictly metaphysical issues, such as the problem of persistence conditions, should be dealt with in isolation from its possible relevance to practical concerns. (see, e.g., Olson 1997: 42, 69)

In this article, without attempting to resolve these perennial issues, I will discuss the relevance of agency for personal identity. Although this article has an introductory form, I offer an opinionated overview of the psychological approach to personal identity, most famously expounded by Derek Parfit (1984, 1995), and the role that agency might play in it.

In the next section, I will provide the background relevant for discerning the importance of agency for personal identity. This will involve introducing the psychologically based criteria of persistence conditions. In the third section, I will introduce and discuss Parfit's Reductionist View of personal identity. In the following section, I will discuss the implications of the Reductionist View for our practical concerns. The last section discusses an agency-based view of personal identity and its prospects for vindicating practical concerns that we relate to the notion of personal identity.

2. Psychological Accounts of Personal Identity and What Matters

There seems to be two dominant positions on the persistence conditions of a person. (Schechtman 2005: 1) One involves the biological view of personal identity that states that persons are essentially human animals or organisms. According to this view, a person at t is the same being at t_1 *iff* the being at t_1 and the person at t are biologically continuous. On this view, a person cannot survive the death of her biological body. According to the so-called neo-Lockean psychological continuity theories, persons are not identical to the biological organisms that may or may not constitute them. Rather, the identity of a person is determined by the psychological relations that comprise one's mental life. On this view, a person could survive the death of her biological body, by, for instance, transferring her memories and experiences into some other functioning body.

If agency has any role in thinking about personal identity, it is likely that this role will be more prominent in psychological accounts of the personal identity. This connection has an intuitive support. To be an agent is to have goals and beliefs about how to satisfy those goals. Furthermore, to be a human agent is to be active with respect to those mental states, to evaluate them, make decisions, and act upon those decisions. Hence, it seems that

agency essentially involves certain psychological features, which enable one to be active with respect to oneself and the world.

In addition, psychological accounts are usually supported by noting their connections with practical concerns that often motivate our interest in problems of personal identity in the first place. (cf. Schechtman 2005) For example, if we imagine that my consciousness and experience will tomorrow be transferred to some other body that will be tortured, then it seems I should be scared and anxious about this future event. This seems to be the case because our intuitions seem to follow the psychological criteria of persistence, and not the fact that we are constituted by some particular organism. (For an alternative view, see Shoemaker 2016, Williams 1970.) Similarly, to use Locke's classical example, if a cobbler and a prince switch bodies, so that prince's consciousness is transferred to the cobbler's body and *vice versa*, it seems that our intuitive response is that now the prince inhabits the cobbler's body and *vice versa*. And for whatever actions we used to hold the prince responsible now we will hold responsible the person who currently inhabits the cobbler's body. (Locke 1690/1998, II.15: 440-441)

In what follows, I will introduce the basic features of the psychological accounts and on this background, in the following sections, explain how agency might be relevant for our conceptions of personal identity.

Contemporary psychological accounts of personal identity draw from ideas developed by John Locke. (1690/1998, §27) Locke famously argued that the idea of a man has different identity conditions from the idea of a person.[1] This can be seen by checking our intuitive responses to hypothetical cases. For instance, let us imagine that some criminal mastermind, doctor *X*, has the technology to change bodies so that he escapes capture after he commits a crime. After performing some evil deed, doctor *X* captures an innocent person *Y* and by using his technology switches their bodies. After every switch, the criminal kills the innocent victim who is now in his former body. The question is, who would we hold responsible for doctor *X*'s misdeeds and would we want to punish the person who is now in *Y*'s body? It seems that we would. After all, the person in *Y*'s body is someone who has consciousness of doctor *X*, identifies with his history and mental states, and continues to carry out his plans. Thus, given our practical concerns we can discern what type of identity our judgments about persons follow and see that what defines personal identity is to a significant extent shaped by our value judgments.

---

[1] In the contemporary terminology, instead of using the word "idea" we would be talking about the concept of a person or a man.

Derek Parfit, in his seminal book *Reasons and Persons* (1984), developed a psychological account of personal identity and argued that what actually matters for our practical concerns is not personal identity but something close to it. To understand his argument, we first need to lay down the basic structure of the psychological view of personal identity.

Parfit develops a neo-Lockean account of personal identity. According to him, the psychological criterion of personal identity over time includes the following:

(1) There is psychological continuity if and only if there are overlapping chains of strong connectedness. X today is one and the same person as Y at some past time if and only if
(2) X is psychologically continuous with Y,
(3) this continuity has the right kind of cause, and
(4) there does not exist a different person who is also psychologically continuous with Y.
(5) Personal identity over time just consists in the holding of facts like (2) to (4). (Parfit 1984: 207)

To understand what this account amounts to, we need to explain points (1)-(5).

Points (1) and (2) can be explained together. Parfit defines psychological continuity in terms of overlapping chains of strong connectedness. He defines psychological connectedness as "the holding of particular direct psychological connections." (Parfit 1984: 206) For example, one of the important psychological relations for personal identity is the continuity of memory. Intuitively, we think that to be the same person from t to t_1 is to be aware or conscious of yourself as the same person. One of the capacities that enables us to be self-aware is memory. However, we can forget things, even important things about ourselves, without ceasing to be the same person. For instance, I may forget what I was doing and experiencing on my third birthday. Nevertheless, it is plausible to think that I am the same person as my three-year-old self.

In order to remedy this problem for the memory criterion, Parfit distinguishes between *direct memory connections* and *the continuity of memory*. A person has direct memory connections with her past self *iff* she can now remember that she had particular experiences in the past. For instance, I remember that this morning I was the person who drank coffee in my kitchen, thus I have a direct memory connection with that person. Furthermore, a person does not have to have direct memory connections with somebody for her to be that person. For *X* to be the same person as *Y* there needs to be *continuity of memory* between X and *Y*. According to Parfit, continuity of memory consists in having an overlapping chain of direct

memories. (Parfit 1984: 205) *X* does not have to directly remember the experiences *Y* had twenty years ago in order for *X* to be *Y*. It is enough that there is an overlapping chain of direct memories.

Parfit generalizes the concepts of direct connections and continuity from the memory condition and introduces other psychological relations that might be relevant for determining the persistence conditions of a person. Importantly, Parfit, at least implicitly, recognizes some basic contours of *agency* as being relevant for personal identity. For instance, he mentions beliefs and desires that one can have over time. He also mentions intentions and later acts that serve as executions of those intentions. (Parfit 1984: 205) Intentions are very important for psychological continuity because their typical structure involves diachronic aspects.[2] At time *t* we may adopt an intention, which structures our plans, beliefs, desires, and characters, and thus governs our behavior until the execution of an action at some later time *t1*.

Including agency as one component of Parfit's psychological criterion might be controversial. Most authors do not regard Parfit as including agency into his psychological criterion of personal identity. For instance, Korsgaard seems to criticize Parfit for not giving an appropriate role to agency in his arguments. (Korsgaard 1989) This might be because Parfit does not really put emphasis on agency. In addition, he discusses the issue of personal identity by putting emphasis on the third personal point of view of agents and their mental states, which might seem to downplay the importance of the active role these mental states might play in personal identity. Nevertheless, as already noted, Parfit's view seems to encompass the agential elements as well. To see this, consider the standard causal theory of action. According to this view

> the agent performs an action only if an appropriate internal state of the agent causes a particular result in a certain way. (…) You turned on the light only if the light came on as a result of some neural and/or mental state you were in. (…) Your action was intentional only if the initiating cause was the desire or intention to turn on the light. If you turned on the light unintentionally, then the light came on because you wanted to do something else instead, such as turn on the fan. So the causal theory says that whether an action was intentional depends on whether it was caused by a particular internal state, a desire or intention to perform that action. (Davis 2010: 32-33, italics in the original)

Thus, according to the standard theory of action, it is necessary to have specific mental states, such as intentions that cause actions in the appropri-

---

[2] For a discussion of the diachronic aspects of intentions and their relevance for personal identity, see (Bratman 2007, essay 2).

ate way, for an action to count as intentional.[3] Different types of intentional actions might require sufficient conditions that go beyond the causal conditions. For instance, to capture full-blown autonomous agency, we might need to refer to additional higher order mental states, such as second order desires or planning attitudes that regulate first order attitudes. (see Davis 2010: 34) Given the standard theory of action, Parfit's psychological criterion seems to satisfy the necessary condition for including agency into considerations that play a role in determining personal identity. In addition, it seems plausible that Parfit's account, with some caveats, could be extended to include also more sophisticated forms of agency. I discuss this option in section 5.

To recap, psychological continuity consists in these overlapping chains of strongly direct psychological connections, including agential components – which Parfit calls strong psychological connectedness.

According to Parfit (1984: 206), psychological connectedness comes in degrees. For instance, X and Y may have many thousands of psychological connections or even just a single connection. X may share with Y all beliefs, desires, and memories and the same intentions and long term plans. Plausibly, in this case X would be the same person as Y. However, if Y had an intention to do something and X remembers having the same intention as Y, and that is the only psychological similarity between X and Y, then presumably they are different people. Thus, Parfit maintains that strong psychological connectedness involves having *enough* of direct psychological connections.

> For X and Y to be the same person, there must be over every day enough direct psychological connections. Since connectedness is a matter of degree, we cannot plausibly define precisely what counts as enough. But we can claim that there is enough connectedness if the number of connections, over any day, is at least half the number of direct connections that hold, over every day, in the lives of nearly every actual person. (Parfit 1984: 206)

This inability to provide a non-arbitrary criterion for strong psychological connectedness will have important consequences for Parfit's claim that personal identity is not what practically matters to us. But before we get to that, we need to further explain Parfit's account of personal identity.

Point (3) states that the psychological continuity needs to be caused in the right way. Causal condition in psychological accounts is usually introduced to handle some problematic cases. (cf. Shoemaker 1984) For instance, one of the earliest objections against using memory as a condition for personal identity is the problem of circularity. For a memory to be part

---

[3] The casual theory of action is not without its critics. For a discussion of some of the objections and a defense of the causal theory of action, see Schlosser (2011).

of the ground of one's persistence through time, the memory needs to be of something that really happened to the person. If I only have memories of events that did not happen, then those memories cannot constitute *my* persistence conditions. Since those things did not happen, they cannot constitute me as the person that I am. Thus, any account of personal identity that uses a memory criterion needs to be able to distinguish between real memories and apparent memories or delusions. However, this seems to involve a circularity. Schechtman gives a succinct exposition of the problem:

> the difference is just that genuine memories are of an experience the rememberer actually had, while delusions are apparent memories of an experience that was not had by the person seeming to remember. Since the memory criterion must define identity in terms of real memories, and real memories are defined in terms of personal identity, the criterion ultimately defines identity in terms of itself. (Schechtman 2014: 22)

It seems, thus, that to determine whether a memory is veridical, we need to know whether the event in question happened to the person. But if this is the case, then we already need to know the identity of the person in order to determine whether she had the relevant experience or not.

The standard solution to this problem involves replacing the concept of memory with a more general concept of quasi-memory. (Olson 2002, §4) Quasi-memory is a memory type experience that is caused in a right way and does not presuppose personal identity. Parfit defines the notion of having an accurate quasi-memory as follows:

> A. I seem to remember having an experience,
> B. someone did have this experience, and
> C. my apparent memory is causally dependent, in the right kind of way, on that past experience. (Parfit 1984: 220)

From having a quasi-memory of cooking lunch today, I cannot conclude that it was *me* who cooked the lunch, rather I can conclude that *somebody* cooked the lunch today. (For further discussion of quasy-memory, see Roache 2006.)

In principle, I can have quasi-memories of somebody else's experience. For instance, we can imagine that a genius neuroscientist can reconfigure my brain states in such a way that they become isomorphic to the relevant brain states of my wife. Now I seem to remember going on a business trip to Brazil and having an awful time there, when in fact it was my wife who went to Brazil and had unpleasant experiences. This possibility aside, quasi-memory, with its causal condition enables us to avoid the circularity objection. The important thing for psychological accounts is that ordinary memory can be seen as a subset of quasi-memories, namely as "quasi-memories of our own past experiences." (Parfit 1984: 220)

It has to be remembered that in psychological accounts of personal identity quasi-memory only partly constitutes personal identity. Other criteria include the continuity of beliefs, desires, values, intentions, etc., which also should be caused in the right way. In particular, Parfit (1984: 207) includes aspects of agency as important elements that need to be properly causally integrated. At one point, he mentions that changes in character need to be caused in the right way if they are going to count as relevant for determining one's personal identity. (Parfit 1984: 207) Continuity of character is preserved if the changes in the character are caused by deliberate decisions, growing older or responses to particular life experiences. However, the continuity of character is hampered if the changes are "produced by abnormal interference, such as direct tampering with the brain." (Parfit 1984: 207) These remarks are congenial to Christine Korsgaard's idea that the right kind of cause is the one that stems from the decisions and commitments endorsed by an agent. (Korsgaard 1989) According to Korsgaard, the psychological connections and their continuity are identity preserving if they are based on *authorial* causes, that is, causes that are based on agent's deliberations and decisions.

Parfit (1984: 207-209) offers different views on how (3) could be understood. On the narrow reading, the *right* kind of cause is the *normal* cause. The normal cause of our psychological continuity could be the brain that implements those connections. In this case, the sameness of the brain would guarantee that the psychological continuity is caused in the right way.

In addition, Parfit (1984: 207-208) distinguishes between the wide and the widest reading of (3). According to the former, any *reliable* cause of psychological continuity is sufficient to be the right cause. On the latter reading, *any* cause can be the right kind of cause for psychological continuity. He does not say much about the difference between the wide and the widest reading of the "right cause," nevertheless he argues that they have better theoretical standing than the narrow reading.

This can be shown by an analogy. Suppose that our brain cells are degenerating and that we have the technology to replace them with some synthetic material. Once all of the organic brain cells die, we still have a functioning brain that causes the psychological continuity since now the synthetic material has taken over the function of the brain cells. If the normal cause were needed for psychological continuity, then in this case after the replacement of brain cells, we would not have the same person, since supposedly this synthetic material is not what *normally* causes the psychological continuity in a person. Based on a similar case, Parfit (1984:

208-209) contends that it is better for a psychological account of personal identity to allow for a wider construal of the notion of the *right cause*.[4]

Clause (4) states that uniqueness is necessary for identity through time. Consider an example. In the so-called simple teletransportation case, we are imagining a person who decides to travel to Mars by means of tele-transportation. This is a type of teleportation familiar from TV shows, such as *Star Trek*. The important difference is that in the simple teletransportation case, when a person enters the machine, her brain gets scanned and the psychological states are copied while her body is destroyed on Earth and rebuilt from organic materials on Mars. Also, the important things that get copied are the brain patterns and psychological connections. Intuitively, it seems that if X decides to travel by teletransportation then the person that is psychologically continuous with X will really be X and not just her replica.

Now consider a case of teletransportation in which X enters the machine, his brain gets scanned and the psychological information is sent to Mars where a new body that is psychologically continuous with X is formed. But in this case the machine does not destroy X's body on Earth. It seems obvious that the person on Mars is not identical with X, even though she is a complete psychological replica of X. It is unclear what this imaginary case shows. It could be taken to show that not even in the simple teletransportation the person on Mars is identical with X. On the other hand, if we follow the intuition supporting the psychological accounts, we can say that in the first case, X is the same person who travels from Earth to Mars *via* teletransportation, while in the second, this is not the case. What explains the difference is the fact that the uniqueness condition is not satisfied in the second case, while it is in the first one.

Clause (5) expresses what Parfit calls reductionism about personal identity. According to him, the only plausible account of personal identity, whether of physical or psychological kind, entails reductionism, in the sense that what personal identity consists in is exhausted in conditions (1)-(4). Since Parfit grounds his conclusion that personal identity is not what is practically important on the idea of reductionism, in the next section I will explain how reductionism is supposed to support this implication.

---

[4] Given Parfit's claim that even the widest reading of the notion of *right cause* can be legitimately adopted, he seems to be trivializing the requirement of having the right cause. (Schechtman 2014: 25, fn. 25) Claiming that *any* cause can be the right cause seems to strip the requirement of any importance for determining personal identity. As already indicated, agency-based accounts limit the right causes to the deliberative capacities of an agent, and to capacities that play a role in unifying and integrating different temporal phases of an agent. See also the discussion in section 5 below.

3. Reductionism and the Non-Importance of Personal Identity

According to Parfit, the presented account of identity is reductionist because it claims that the identity of a person consists in obtaining of conditions (1) to (4). In fact, there is no need to invoke any *further* facts, beyond those that underlie psychological continuity that is caused in the right way. According to Parfit, the Reductionist View includes the following two conditions:

a. that the fact of a person's identity over time just consists in the holding of certain more particular facts, and
b. that these facts can be described without either presupposing the identity of this person, or explicitly claiming that the experiences in this person's life are had by this person, or even explicitly claiming that this person exists. These facts can be described in an impersonal way. (Parfit 1984: 210)

In the case of psychological theories of identity, condition (a) states that identity consists in and is exhausted by the psychological continuity that is caused in the right way. Condition (b) states that facts about personal identity can be described in such a way that does not presuppose the identity of a person and by using a language that is impersonal, or we might say, based on a third-person perspective.

Parfit opposes reductionism to what he calls the *Further Fact View* (Parfit 1984: 210) Non-reductionists claim that psychological continuity does not exhaust facts about personal identity. Reductionists deny this, and claim that there is no further fact that determines personal identity. According to Parfit, a good test for whether someone's view of personal identity is reductionist is the following:

> If we accept a Reductionist View, there may be cases where we believe the identity of such a thing to be, in a quite unpuzzling way, indeterminate. We would not believe this if we reject the Reductionist View about this kind of thing. Consider, for example, clubs. Suppose that a certain club exists for several years, holding regular meetings. The meetings then cease. Some years later, some of the members of this club form a club with the same name, and the same rules. We ask: "Have these people reconvened the very same club? Or have they merely started up another club, which is exactly similar?" There might be an answer to this question. The original club might have had a rule explaining how, after such a period of non-existence, it could be reconvened. Or it might have had a rule preventing this. But suppose that there is no such rule, and no legal facts, supporting either answer to our question. And suppose that the people involved, if they asked our question, would not give it an answer. There would then be no answer to our question. The claim "This is the same club" would be neither true nor false. (Parfit 1984: 213)

According to Parfit, in the case of a club, there is a possibility of indeterminacy of identity because the identity of a club does not consist in anything beyond the existence of its members and their behavior and attitudes. Thus, when we know everything about club members and their relations, there is *no further fact* that might determine whether one club is the same club at time *t* and *t1*. If facts about people that comprise a club cannot decide the issue, then the identity of the club is indeterminate.

When the question whether a thing at *t* is the same as the thing at *t1* is indeterminate in the above sense, then we can conventionally decide to say that at *t* and *t1* we have the same thing or we can say that we have a different, albeit, very similar thing. For example, we can either say that the club at t is identical to the club at *t1* or we can say that the two are similar in all respects but are not actually identical. According to Parfit, when the answer is conventional in this sense, then, although the question whether *X* is identical to *Y* is sensible, it is actually *empty*. (Parfit 1984: 213)

Once we accept that "person" does not refer to a separately existing entity and does not involve further facts that go beyond those involved in psychological continuity and/or psychological connectedness and their physical implementation, according to Parfit,[5] we get to two different, albeit, related conclusions. One is the previously mentioned claim that reductionism is committed to the possibility of indeterminism about identity. The second is that identity is not what really matters in survival.

To see why reductionism is committed to indeterminism we have to remember that, according to the psychological criterion, personal identity reduces to a *unique* psychological continuity, which consists in overlapping chains of psychological connections with the right kind of cause. Since there is no non-arbitrary way to say to what degree the overlapping chains of psychological connections need to obtain for numerical identity through time, we can imagine that degrees of connectedness fall on a spectrum. (see Parfit 1995) On the one side of the spectrum there will be a complete psychological continuity between *X* and *Y* from *t* to *t1*. On the other side of the spectrum there will be no psychological continuity between *X* and *Y*. In the first case *X* and *Y* would be the same person, while in the second they would be different persons. Now suppose that there is a

---

[5] Parfit extensively argues for the Reductionist View and the view according to which a person is not some separately existing entity (Parfit 1984, sec. 88; see, also, Parfit 1995). Without going into details, here I just note that Parfit persuasively argues that the only viable alternative to reductionism is some form of mind/body dualism. As is well known, there are great difficulties for defenders of dualism to explain the identity conditions of a separately existing self, and in what way persons could exist as non-physical substances (see, e.g., Maslin 2001: §1).

neurosurgeon who can change the degree of psychological connectedness (memories, plans, character, etc.) between X now and X at some later time by changing the neural activation patterns in X's brain. Small changes in psychological connectedness between X at t and $t1$ will not affect his personal identity. However, if little changes are progressively made, at some point the degree of connectedness will fall somewhere in the middle of the spectrum. When that happens, we will not be able to say whether X will continue to be the same person or an additional small change will make X disappear.

The case is similar to the *Sorites paradox*. Adding one more grain of sand will not make a heap. If we add enough grains to one pile, it will make a heap, but there is no non-arbitrary way of telling when a non-heap becomes a heap. A similar thing seems to follow about personal identity if reductionism is true. It is possible that we know all the identity relevant facts, but still we cannot say whether some person continues to exist or ceases to exist. That is why Parfit says that there are *empty* questions about personal identity. Even if we are looking at the middle of the spectrum we can ask whether X still exists or ceases to exist. We can even decide to say that if there is a 60% of overlapping chains of psychological connections between X at t and Y at $t1$ then X is identical with Y. Nevertheless, since, by the supposition of reductionism, we already know everything that is to know about psychological facts and degrees of connectedness, this cut off line would be stipulated and not imposed on the psychological criterion by the facts that determine personal identity. If we care to provide an answer to questions of identity in these borderline cases, the answer could only be given by *fiat* and not determined by metaphysical facts alone.

The second implication of reductionism is that personal identity is not what matters in survival, rather, what is important is psychological continuity and/or connectedness. Parfit refers to psychological continuity and/or connectedness with the right kind of cause as the Relation R. (Parfit 1984: 215) To see why according to reductionism Relation R is more important than personal identity let us consider another example.[6] Normal human brain has two hemispheres that are connected with the brain area called *corpus callosum*. Let us imagine that both hemispheres could implement in its entirety Relation R. In this imaginary case, if X's left hemisphere were to be damaged, the right hemisphere could take over and X's personal identity would be preserved. The same thing would happen if the right hemisphere were to be damaged. Now suppose that X's body and the

---

[6] We have to bear in mind that according to the psychological criterion personal identity consists in R and the uniqueness condition.

right brain hemisphere are destroyed and a doctor decides to transplant his left-brain hemisphere to another person's body that is currently brain dead. In this case, we are inclined to think that *X* would continue to exist in another person's body. Now imagine that for some reason *X*'s both brain hemispheres are not damaged but they are separately transplanted into two different bodies, *Y* and *Z*. Since, by hypothesis, the two brain hemispheres are both fully functional and can fully implement and preserve *X*'s psychological connections, the bodies *Y* and *Z* both have identical *R*'s to *X*. Furthermore, since identity is a transitive relation and *Y* and *Z* are not numerically identical persons, they cannot both be identical to *X*. In fact, according to the Reductionist View, neither is identical to *X* since they both stand in the same relation *R* to *X*, and thus the uniqueness condition is not satisfied.

Nevertheless, it would be inappropriate to say that *X* dies or ceases to exist once his brain hemispheres are divided. As Parfit writes,

> We might say: "You will lose your identity. But there are at least two ways of doing this. Dying is one, dividing is another. To regard these as the same is to confuse two with zero. Double survival is not the same as ordinary survival. But this does not make it death. It is further away from death than ordinary survival." (Parfit 1984: 262)

It seems that this example shows that what really matters in continued existence is not personal identity, rather it is Relation *R*. In cases of division, where Relation *R* takes a branching form, even though we cannot say that literal personal identity is preserved, people do not just cease existing. As Parfit writes, two is different from zero. In ordinary life, Relation *R* tends to coincide with personal identity, but in these extreme cases we see that they can come apart. And when they do, it seems that we are inclined to believe that what really matters to us is that at least one person will bear Relation *R* to us.

To support the latter claim consider the following example. At time *t* I find out that at some future *tn* I will divide into two people who will have identical Relations R to me. In that case, would I stop thinking about the future and decide to live like *tn* will be my last day alive or would I be preparing for the future and devise plans that could be more easily executed by my branching counterparts? For example, I might decide to write a book and to ease the coordination problems, I might indicate which of the two of my continuers should devote his time to that activity. The other one could then devote his time to accomplish some of my other deeply held desires. Even though there might not be anything essentially different between holding an attitude as in the first (where a person thinks his life

is finished) and the second example, it is plausible that we would be more inclined to have the latter attitude towards our knowledge that in the future we will be divided.

4. Problems with the Reductionist View: The Extreme Claim

The importance of agency and related considerations for thinking about personal identity can be discerned once we take into consideration a common objection against the Reductionist View. I will turn to the agency based solution in the next section. In what follows I will present the problem and define the notions that underlie it.

According to what Parfit calls the Extreme Claim, if the Reductionist View is true then it could be argued that we would not have any special reason to care about our own future. We would have equally valid reason to care about ourselves as much as anybody else, we should be indifferent to whether we live or die, etc. (Parfit 1984: 306-307) Many authors find these ideas problematic because they think that many of the practical concerns that we have and care about could not be explained or justified without some reference to the notion of personal identity.

The objection is that if reductionism is true then we could not make sense of our prudential, ethical, and "forensic" practices. Intuitively we think that we bear special relations to our past and future selves. As Korsgaard writes, "I am responsible for my past self, and I bear the guilt for her crimes and the obligations created by her promises. I am responsible to my future self, for whose happiness, since it will one day be mine, it is rational for me to provide." (Korsgaard 1989: 108)

One version of an objection based on the Extreme Claim can be formulated as follows:

I. If reductionism about personal identity is true, then it would not make sense to hold people responsible for their past actions. (One version of the Extreme Claim)

II. Holding people responsible for their past actions is an important part of our social and ethical practices. Thus, it makes sense to hold people responsible for their past actions.

Therefore,

III. Reductionism about personal identity is incompatible with our social and ethical practices.[7] (see Schechtman 2014: 34-35)

---

[7] Some people would not have problems with accepting the first premise of this argument and therefore with accepting the conclusion. These people would be characterized as revisionist with respect to our common social and ethical practices. Parfit himself seems to be inclined to accept many of the consequences of the Extreme Claim. (see Parfit 1984: ch.14)

The problem with (III), for instance, is that, if true, nobody could deserve to be punished for their past crimes.

Support for (I) comes from the so-called fission cases. Let us suppose that I am a hardened criminal whose two brain hemispheres were for some reason transplanted into two different bodies, *Y* and *Z*. Intuitively, a person should be punished for a crime only if she committed that crime. According to the Reductionist View, there is no non-arbitrary reason to say that I am identical with either *Y* or *Z*. Since *Y* and *Z* are different people, it is plausible to say that I am neither *Y* nor *Z*. Therefore, since there is no *me* after the transplantation, there is no person that should be punished for my crimes. According to this argument, accepting the Reductionist View would commit one to a revisionist perspective on our social and ethical practices.[8]

One could object to the argument that the first premise is not true. Indeed, in the fission case we could not blame *X* for his misdeeds, because after the fission there is no *X*. But fission cases represent only abnormal situations. Normally, we will have a person that does not share with other selves, to a significant degree, Relation *R*. Thus, normally, personal identity would be important for ascribing responsibility and other practical notions. However, this objection would arguably miss the point. The proponent of the argument could say that what it really shows is that when we think about responsibility, for instance, what is really important is Relation *R*. If we are willing to punish *Y* and *Z* for *X*'s crimes, then what really matters is not *X*'s identity, but rather the inheritance of his *R*. Similar considerations could then be applied to normal, non-fission cases.

A more promising objection is that the Reductionist View does not entail some unpalatable version of the Extreme Claim. At least, this is not shown by relying on the aforementioned fission case. In fact, it is plausible that our moral practices only require the continuation of Relation *R*, and not the satisfaction of the uniqueness condition. (Bělohrad 2014a: 316-317; see Parfit 1984: ch.13) In the fission case, it does not offend our intuition to say that *Y* and *Z* should both be punished for *X*'s crimes. After all, they directly inherit what our responsibility judgments seem to track, namely her personality and other components of Relation *R*. Thus, we might say that our practical judgments track, what we might call, *moral selves* and not personal identities *per se*.[9] The purported difference between a moral self

---

[8] Here the supposition is that even after the fission we would still want to punish *X*'s descendants.

[9] For a similar claim that the relation between responsibility and personal identity should be loosened. (Beck 2015: 315-316) The notion of moral self that I use here is similar to the one Marya Schechtman uses in her book *Staying Alive*. However, it should

and the self of personal identity consists in the fact that while the latter is unique to a person, the former comprises set of mental states, personality traits, dispositions, and a history that, in principle, might be shared by different persons.

The idea of a moral self, I believe, captures the intuitions underlying our practical judgments in the fission cases. For instance, let us imagine that a criminal X decides to split her two brain hemispheres and implant them into two different bodies (Y and Z) in order to increase her chances of escaping the police. Let us say that her reasoning is that if she divides herself into two people who will inherit her mental life, memories, plans, projects, etc., there is a greater chance of one of them escaping. Now, neither Y nor Z is strictly identical with X. They are two different spatiotemporally individuated objects. However, they are the same person as X in the sense that they are psychologically identical with X. They committed the same crimes, they have the same history, memories, plans, dispositions, etc. In other words, we can say they have the same moral character or the self. Because of that moral self, we feel that Y and Z are responsible for the crimes X committed in the past and they should be punished for them. Thus, this example seems to indicate that our responsibility judgments track what we might call our *moral selves*, and not the strict identity of a person.

However, reductionism about personal identity has other seemingly undesirable consequences. If that which really matters in our practical concerns is our moral self, which is underpinned by Relation R, and not our uniqueness (strict personal identity), then it seems we would not have a reason to be especially concerned about ourselves. This point is nicely illustrated by Radim Bělohrad:

> Another practical consequence of reductionism results from the fact that part of relation R, namely connectedness, holds in degrees. That is, I may be more or less connected to my future and past selves. Thus, if R justifies attributions of responsibility, I may be less responsible for the actions of my distant past self than for my yesterday's self. Similarly, if R is what justifies the rationality of my concern for the future inhabitant of my body, when

---

not be confused with it. Schechtman (2014: 14-15) distinguishes *moral selves* from *forensic units*. She seems to think that our moral practices and practical concerns presuppose the existence of forensic units and not just moral selves. However, the distinction between moral selves and forensic units is subtle. (see, also, Bělohrad 2014b: 566-567) Forensic units are entities that exist in virtue of being the proper targets of our practical concerns (such as, blame, self-concern, etc). Moral self, on the other hand, is comprised of contingent properties that give content to a forensic unit. In other words, having a moral self determines what practical judgments *actually* apply to individual forensic units or persons. Thus, forensic units provide a prerequisite for the existence of moral selves.

> R holds to a low degree, so should my concern. This aspect, in turn, leads to an increase in the plausibility of paternalism, because great imprudence with respect to my distant future self is seen as violating my obligations to others, rather than myself, thus becoming immoral, rather than irrational. (Bělohrad 2014a: 317)

In effect, the unimportance of personal identity leads to the unintuitive view that we do not have rationally binding reasons to be concerned about our distant future selves. Since Relation *R* underpins personal concerns, the more different (in terms of *R*) we are from our future self the less reason we have to be rationally concerned about her. In other words, our distant future selves are to us like any other distant person we may or may not know. However, this seems unintuitive. It seems reasonable to say that we have more reason to care about that person who will inhabit our body in the future than somebody who is physically completely distinct from us. Thus, it is not clear how persuasive this example is.

One could argue that the Reductionist View cannot be true since psychological connections do not support the unity that is presupposed in personal identity. The unit to which we apply practical judgments, such as judgments of responsibility and blame, presuppose that there is a deep unity of *consciousness*. This view has roots in Locke who maintained that personal identity pertains to the unity of consciousness and not to the existence of bodily substances. (Locke 1690/1998: II.9) If the Reductionist View were false, then the Extreme Claim would lose its support. However, Parfit has an argument that pertains to show that personal identity does not presuppose the deep unity of consciousness.

To illustrate this claim, we can once again examine one of Parfit's examples. (see Parfit 1984: 246-247) Let us suppose that my two brain hemispheres possess the same abilities and each can function as a separate unit that can implement conscious experiences and cognitive functions. In addition, suppose that I am able to disconnect the communication between the hemispheres, so that they operate as separate cognitive units. Since both hemispheres can support consciousness, when I disconnect them it is as if I have two minds. This ability would allow me to solve some tasks more effectively. For example, I might be taking a physics exam and not be sure how to answer a question. I see at least two ways in which the question could be answered. Since the time is pressing, I decide to disconnect my brain hemispheres and let each try out one possible solution to the problem. During that time, there is no unity of consciousness, each brain hemisphere is conscious of what it is doing and what it has control over.[10]

---

[10] For instance, left brain hemisphere controls the right arm, has a visual field of the right eye, etc. The opposite could be true of the right brain hemisphere.

After some time, when the two hemispheres do their parts, I reconnect them, that is, reunite my mind (consciousness), and compare the solutions. Furthermore, after the unification, I remember everything that each hemisphere did separately, and in that sense I am psychologically continuous with both consciousnesses that existed for a while.

This example is supposed to show that "a person's mental history need not be like a canal, with only one channel, but could be like a river, occasionally having separate streams." (Parfit 1984: 247) According to Parfit, although portraying a surprising possibility, the coherence of this example shows that I could be the same person from t to tn, without exhibiting the unity of consciousness. Thus, we could conclude that the unity of consciousness is not necessary for personal identity.

Other authors have tried to resist the Extreme Claim by giving an account of personal identity that can vindicate our intuitive practical concerns. Most notably, Christine Korsgaard (1989) argued that an agency-based view could vindicate the importance of the deep unity that personal identity presupposes in our practical concerns. She seems to agree that if reductionism about personal identity is true then we are committed to the Extreme Claim.[11] Thus, she develops an agency-based view of personal identity that is supposed to show why and in what way the reductionism about personal identity is not true. In the next section, I will present Christine Korsgaard's (1989) and Michael Bratman's (2007) views on the role of agency in personal identity. I will argue that agency-based accounts are in principle compatible with the Reductionist View. Then I will evaluate the prospects of the agency-based view for solving problems related to the Reductionist View.

5. Reductionism and the Agency-Based View of Personal Identity

In her seminal paper, Korsgaard (1989) argued that Parfit's psychological criterion of personal identity is too theoretical, and when we direct our attention to a practical perspective, we will see that there is stronger unity of the self than Parfit envisioned. I will argue that Korsgaard's agency-based approach can be seen as an extension of Parfit's Reductionist View.

What creates the appearance that Reductionist views cannot account for deeper unities that comprise a person's self is Parfit's theoretical perspective. The most important element of the psychological criterion is the psychological continuity and/or connectedness with the right cause. When

---

[11] This seems to be the standard understanding of Korsgaard's position (see, e.g., Schechtman 2008: 407-408).

Parfit discusses this criterion, one is left with an impression that in a standard case psychological continuity is a collection of very loose relations. Given that the Reductionist View implies the possibility of indeterminacy of identity, this impression might seem to be justified. In addition, this impression is reinforced by Parfit's claim according to which any cause can count as the right cause for maintaining personal identity (see, above, ft. 4).

However, the possibility of indeterminacy does not imply that in a standard situation a person's identity will be indeterminate. Korsgaard (1989) argues that when we turn from Parfit's theoretical or metaphysical perspective to a more practical or decision-making perspective, we will see that a deeper unity underlies psychological connections that define a person's identity. In fact, Korsgaard, in a Kantian fashion, argues that from the first person perspective, that is, from the perspective of agency we are *forced* to postulate a locus that unifies and sustains, in a right way, all the psychological relations that determine one's personal identity. The relation between the first-person perspective and agency might not be straightforward. I will shortly discuss this issue in order to show in what way they are noncontroversially related.

Lynne R. Baker (2011) provides a simple argument for the view that being an agent involves having a first person perspective on things. Let us recall that on a standard theory of action an action is intentional only if it is produced, in an appropriate way, from beliefs, desires, or intentions. Thus, actions are events that can be explained by invoking the relevant mental states. The fact that these mental states can be used for explaining action indicates that they can be used in some forms of instrumental or practical means-end reasoning processes that lead from those mental states to the performance of the action. Thus, if agents perform their actions in virtue of mental states they possess, and those states may figure in practical reasoning, then an agent is someone who has an *ability* to engage in at least some forms of instrumental or practical means-end reasoning.[12] However, explaining someone's action normally involves seeing things from her own perspective. Donald Davidson has made this point especially salient a long time ago:

> A reason rationalizes an action only if it leads us to see something the agent saw, or thought he saw, in his action-some feature, consequence, or aspect of the action the agent wanted, desired, prized, held dear, thought dutiful, beneficial, obligatory, or agreeable. (Davidson 1967: 685)

---

[12] According to Baker, these considerations indicate that there is a conceptual connection between agency and the ability to engage in primitive forms of instrumental reasoning. (Baker 2011: 3)

When we combine this consideration with the idea that intentional action presupposes having an ability to engage in practical reasoning, we get to the view that "the agent reasons about what to do on the basis of her own first-person point of view. It is the agent's first-person point of view that connects her reasoning to what she actually does." (Baker 2011: 3) Thus, we can conclude that the perspective of practical reasoning involves a first-person perspective. Now we can return to the discussion of Korsgaard's agency-based view of personal identity.

Korsgaard (1989: 109) starts by asking why we have an experience of ourselves as being a unified entity in a particular moment? If we do not believe in a Cartesian Ego, what could explain the experience that we are the person that we are, rather than someone else? Korsgaard suggests that the answer lies in our ability to make decisions and act. When we take a practical point of view, we see that we cannot but to think of ourselves as unified active subjects.

Korsgaard claims that two considerations ground this view. The first "is the raw necessity of eliminating conflict among your various motives." (Korsgaard 1989: 110) The second ground for unity can be discerned from "the unity implicit in the *standpoint* from which you deliberate and choose." (Korsgaard 1989: 111) Let us start by considering the first ground.

There are many moments in life in which we are confronted with different desires, considerations, and options that suggest incompatible courses of action. Since we often successfully act, it means that we often manage to resolve these conflicts. When we resolve them, we act as a unified agent. Korsgaard illustrates this point with the split-brain hemispheres example:

> So imagine that the right and left halves of your brain disagree about what to do. Suppose that they do not try to resolve their differences, but each merely sends motor orders, by way of the nervous system, to your limbs. Since the orders are contradictory, the two halves of your body try to do different things. Unless they can come to an agreement, both hemispheres of your brain are ineffectual. Like parties in Rawls's original position, they must come to a unanimous decision somehow. You are a unified person at any given time because you must act, and you have only one body with which to act. (Korsgaard 1989: 110-111)

The first source of unity comes from the *necessity* to act, which includes resolving conflicts and making decisions as a unified subject. Korsgaard mentions that we have one body, which might be taken as implying that this very fact contributes to the unity of a person. However, having one body is not essential for Korsgaard to make her point. We can confirm this contention by imagining a situation in which we can control two bodies in the same way in which we control two hands. So the pressure to act as a unified agent does not originate from the constraint of a particular body,

rather it seems to be consistent with the idea of agency that is distributed across different bodies.

The second ground involves thinking about the first-person perspective from which we deliberate and make decisions. Korsgaard claims that when we reflect on the deliberative standpoint we are compelled to think of ourselves as unified subjects.

> It may be that what actually happens when you make a choice is that the strongest of your conflicting desires wins. But that is not the way you think of it when you deliberate. When you deliberate, it is as if there were something over and above all your desires, something that is you, and that chooses which one to act on. The idea that you choose among your conflicting desires, rather than just waiting to see which one wins, suggests that you have reasons for or against acting on them. And it is these reasons, rather than the desires themselves, which are expressive of your will. The strength of a desire may be counted by you as a reason for acting on it; but this is different from its simply winning. This means that there is some principle or way of choosing that you regard as expressive of yourself, and that provides reasons that regulate your choices among your desires. (…) This does not require that your agency be located in a separately existing entity or involve a deep metaphysical fact. Instead, it is a practical necessity imposed upon you by the nature of the deliberative standpoint. (Korsgaard 1989: 111)

Here Korsgaard points out that our natures as beings who are faced with choices and can reflect on the mental attitudes and situations we find ourselves in, practically force us to conceptualize ourselves as *unified* sources of agency.

Korsgaard emphasizes the distinction between the theoretical or metaphysical and practical or agential (and thus the first-person) perspective, in a way that might be interpreted as being incompatible with the Reductionist View.[13] For example, the insistence that the unity of a person can be discerned only from an agential point of view, which is naturally interpreted as the first-person point of view, might be taken to negate the Reductionist thesis that personal identity can be determined from an impersonal or third-person point of view (see Korsgaard 1989: 193).[14] Without getting

---

[13] Schechtman (2008) seems to endorse this reading of Korsgaard.

[14] One way in which Korsgaard is opposing Parfit's reductionism is by claiming that a person should be compared to a state and not to a nation or a club. (Korsgaard 1989: 114-115) While nations are just mereological sums of the people that live in a certain territory, states are more than that. On Korsgaard's view, a state "is a moral or formal entity, defined by its constitution and deliberative procedures." (Korsgaard 1989: 114) However, even if the notion of state provides a better analogy, this does not by itself provide a reason to think that reductionism about personal identity is false. To recall, reductionism includes the claim that personal identity consists in holding of some more particular facts (such as, the obtaining of some set of psychological connections) and the possibility of theorizing about personal identity from an impersonal or

into the exegesis of Korsgaard, I will argue that an agency-based view of personal identity might be construed in a way that is compatible with the Reductionist View.

As Korsgaard points out, emphasizing the practical perspective does not involve commitment to some additional metaphysical fact. Thus, taken at face value, it does not contradict the Reductionist View of personal identity. It merely involves a change of perspective, which involves paying closer attention to those aspects that make us agents.

Parfit's discussion of overlapping chains of psychological connections suggests an overly modular view of a person. (see also Bělohrad 2014a: 320-321) However, nothing in his account commits us to this view. Once we direct attention to the structure and organization of psychological connections, a different picture of the Reductionist View will emerge. In fact, when arguing against the narrow construal of "the right kind of cause" Parfit himself recognizes that what is important about psychological continuity is its organization. That is why Korsgaard construes his view as claiming that "persisting identity is simply formal continuity plus uniqueness." (Korsgaard 1989: 106-107)[15] However, I think that already in this formulation there is a gesture towards a psychological criterion that can account for deeper units that underlie personal identity.

David Shoemaker (1996) has argued that Parfit's psychological criterion of personal identity and Korsgaard's agency view of the self are not incompatible, at worst they provide complementary pictures of a person's identity. I concur with this view. What is hidden in Parfit's discussion of the psychological criterion, however, is the emphasis on the structure or the organizational elements that keep together and unify all these different overlapping chains of psychological connections that underpin psychological continuity.

The latter claim can be explained by an analogy with a solution to a general problem from the philosophy of action. Mariam Thalos (2007: 127) indicates that in the Davidsonian tradition in philosophy of action, the relation between mental motivation and action was conceived in quasi-Newtonian terms. The idea is that desires, as paradigmatic pro-attitudes (plus instrumental beliefs), cause action as a function of their strength.

---

third-person point of view. On the face of it, the idea that persons are more like states than nations does not violate the two conditions. Or if it does, this is not obvious and some argument should be provided for that claim.

[15] Korsgaard uses the term "formal" in the Aristotelian sense, where it is contrasted with material. In her words, psychological properties relevant for identity are determined by "the way the matter is organized, not in the particular matter used." (Korsgaard 1989: 106)

From a third-person perspective, this picture looks very passive. It portrays intentional action as being a linear sequence of events. However, this cannot generally be the case. This is especially clear in cases of expert performance. For instance, when experienced drivers drive a car, they usually do not pay attention to all the details that are involved in driving. Steering the wheel and moving one's leg from the throttle onto a brake pedal involves a sequence of movements that need to be properly coordinated in order to translate into successful action. However, there is no need to posit special desires that govern them. The solution is to think about agency as underpinned by hierarchical systems that top out in higher-order cognitive processes that play a role of a controlling device. If things go as planned the agential system runs on an autopilot, but when something goes awry then the controlling device usually takes over. (see Thalos 2007: 132-133)

Similarly, Parfit's psychological criterion of personal identity seemingly portrays agents as inactive bundles of linearly ordered psychological processes. In contrast, Korsgaard's emphasis on the practical perspective indicates that agents are not only passive bundles of unidirectional psychological processes, rather they are active bundles of processes whose activity or the ability to control that activity forms and maintains the unity of the bundle. Moreover, this dynamic picture of the self can also be discerned in Parfit's writings. In *the Nineteen Century Russian* story, Parfit (1984: 327) describes a young Russian who is about to inherit a large amount of land. Since he is an ardent socialist, he decides that once he inherits the land, he will donate it to peasants. However, he is also aware that once this happens in the future he might become a different person, someone who does not have socialist ideals anymore. Since he sees his current ideals as essential to his identity, he decides to do two things:

> He first signs a legal document, which will automatically give away the land, and which can be revoked only with his wife's consent. He then says to his wife, "Promise me that, if I ever change my mind, and ask you to revoke this document, you will not consent." He adds, "I regard my ideals as essential to me. If I lose these ideals, I want you to think that I cease to exist. I want you to regard your husband then, not as me, the man who asks you for this promise, but only as his corrupted later self. Promise me that you would not do what he asks." (Parfit 1984: 327)

With this example, Parfit seems to admit that current commitments, plans, and ideals structure and give contours to a person's self. In this sense, Parfit may also be interpreted as claiming that organizational aspects of agency play a determining role in delineating a person's identity. Thus, what comprises personal identities on this view are the agential structures, which involve desires, beliefs, intentions, values, and plans, that control and coordinate different sequences that compose actions.

In general, the notion of controlling structures can be used to naturalistically ground Korsgaard's ideas about the need to resolve conflicts among motivations and to regard oneself as a separate source of agency. Wayne Christensen (2007) provides an empirically grounded evolutionary story of how organisms when encountering a complex environments benefit greatly from acquiring more and more centralized higher-order controlling structures that can successfully maximize their fitness. Control structures function on the basis of feedback loops that send and receive signals from different components that they govern and respond to. Once those control structures have evolved they determine which options should be pursued, they resolve conflicts produced by subordinated systems, and coordinate activities of these subsystems to act successfully. Importantly, nothing compels us to think that these feedback structures have to be consciously accessible. Nevertheless, from an internal perspective, we can plausibly expect that these control structures, given that they are centralized in higher-order cognitive systems, ground a sense of unity in human agents. In effect, we can say that control structures ground the unity, which provides the locus that is an appropriate target of our practical concerns.

The important difference between the agency-based account that I sketched, and Parfit's account is that Parfit seems to allow any kind of cause to play a role in determining the psychological relations that are relevant for personal identity. (Parfit 1984: 207-208) In the agency-based accounts, the right kind of cause must stem from the capacities that underlie agency. In commonsensical terms, these are the capacities that enable us to act for reasons. Thus, Korsgaard talks about causes that enable "authorial connectedness." Those are exactly the capacities that enable us to make decisions and resolve internal conflicts. Using the present terminology, we might say that the controlling structures, which form a basis of human agency, provide a hierarchical framework that shapes the loci of potential experiences, values, and decisions, which in effect ground the relevant psychological connections.

The agency-based account of the self can still be regarded as reductionist. It does not postulate primitive metaphysical selves. In addition, it preserves the indeterminacy related to the Reductionist View, given that the control structures and the components it governs can be replicated or obliterated in the same sense and to the same degree in which psychological continuity theorists suppose that psychological processes can be replicated or obliterated. In addition, personal identity in this sense can be described from a third person point of view, for instance, by talking about control functions and its effects on the behavior of an agent (see Thalos 2007). Thus, it seems that if the Reductionist View is committed to some

form of the Extreme Claim, then the agency-based view will inherit its negative aspects.

However, the agency-based view enables us to see that the Reductionist View does not have all the negative consequences associated with the Extreme Claim. For instance, the agency-based view can accommodate and explain the fact that we have a special reason to care about our own future selves. This aspect of the agency-based view is especially salient in Michael Bratman's (1999) planning theory of agency.

What is distinctive about human agency is the ability to reflect upon our mental states, make plans that structure and govern our daily activities, and the fact that we conceive our agency as extended through time. (Bratman 2007: 21) For the present purposes, the latter two features are more important. The fact that we form plans and that we conceive our agency as being extended in time, indicates that we have foresight and an ability to control our actions when considering what will happen in closer or farther future. Furthermore, our ability to make short or long-term plans sets functional constraints on our available options. If I decide to go to work every morning, then I will have to settle on a plan that will enable me to successfully execute that decision. For instance, the ensuing intention to go to work will constrain me to wake up every morning at a particular time, to choose the most suitable route and means of transportation, to prepare lunch for that day, and so on and so forth. In addition, I will have to fill in that plan with further subplans that can respond to contingencies that might interfere with smooth execution of my intention to go to work every morning. These subplans might include the problem of deciding whether to cook at home or to buy lunch at work. Similarly to Korsgaard, what distinguishes us from other possible types of agents, according to Bratman, is our ability to conceive ourselves as beings that have this type of temporally extended agency. (Bratman 2007: 29)

What is important in this picture is that it explains why we have a reason to care about what happens to ourselves in the future. Given that we can form plans and that our agency is typically temporally extended, we have a reason now to care about what will happen to us in the proximal or more distant future. In particular, according to the agency-based view, what upholds and determines the psychological connections and continuities relevant for personal identity is to an important extent "a result of the agent's activity." (Bratman 2007: 30) What gives me a reason to care about someone who will inhabit my body in the future is provided by the fact that that someone is psychologically continuous and connected with me in virtue of being constituted and/or supported by my temporally extended agency.

It could be objected that while the agency-based view can explain why we have a reason to be self-concerned in shorter time spans, it will have a hard time explaining why we should care about the temporally more distant inhabitants of our bodies. Bělohrad advances this objection as follows:

> according to Korsgaard, living a life consists in planning and executing projects. It is the projects that force the person's identification, that is, authorial connectedness with a future self. The problem is that people's projects hardly ever span the extent of whole lives. Korsgaard may have shown that in order to carry out a plan, unity is required. But what she has failed to show is that these plans that people have and derive reasons from span their whole lives. (Bělohrad 2014a: 323-324)

Bělohrad, most notably, substantiates this objection by relying on empirical studies that provide evidence that people are relatively poor at long-term planning, delaying gratification, and in general tend to discount the value of future events or options. (Bělohrad 2014a: 325-326)

If agency-based account is relevant for justifying prudential concern, it must be able to provide some response to this objection. Here I will just sketch a possible route that an agency-based theorist might take. Notwithstanding the empirical facts about discounting, it can be replied that self-concern is grounded in the fact that people have *capacities* or *dispositions* for planning and extended agency. However, capacities or dispositions do not have to be manifested on every occasion. In addition, we have the ability to think of ourselves as having the capacity for extended agency. We project ourselves, grounded on our ability for agency, into the future. If we did not have these capacities, we would not find it rational to care about what happens to our future selves. In fact, on the one hand, a plausible explanation of why we think self-concern is rational is exactly the fact that we have a capacity for agency and planning and that we see ourselves as this type of agents. On the other hand, it is plausible that the same facts also explain why we normally think it is irrational to be poor at planning and to discount the future. We might conclude that having the capacity for temporally extended agency is a prerequisite of that aspect of personal identity that underpins the rationality of self-concern.

In this respect, the agency-based account mitigates at least one aspect of the Extreme Claim that is normally associated with the Reductionist View. In addition, the agency-based view can accommodate the idea that consciousness is not essential for personal identity. If we go back to the physics exam, we can see that the person with a divided consciousness is the same person as the one who decides to solve the exam by dividing the consciousness. The decision to solve the task in this way and her capacity to execute that intention through temporally extended agency grounds her identity through time.

However, it seems that the agency-based view does not mitigate all the misgivings related to the Reductionist View. In particular, adopting the agency-based view does not vindicate the idea that personal identity, as opposed to Relation *R*, underlies our most cherished practical judgments. For instance, in the fission case, it seems that our practical judgments track Relation *R* and not personal identity.

That being said, in the remainder of this section, by utilizing the fission case, I will explore how the agency-based view could provide a clue in what way personal identity might be spatially and temporally extended.

We start with the observation that agency seems to underpin the moral self, which, I maintained, is what our practical judgments actually track in the fission case. In order to possess a moral self (described in common-sensical terms), one has to possess capacities that underlie normal human agency. In addition, an agency-based account might explain in which sense our practical judgments might be, after all, tracking personal identity in the fission case. As an intuition pump, imagine that the criminal *X*, in order to enhance his chances of escaping the law, devises a plan which has the following elements: *X* decides to divide his brain hemispheres into two bodies, *Y* and *Z*. In addition, he devises total life-plans for both, *Y* and *Z*, in a way that will enable both of them to always perfectly coordinate their actions in escaping the hand of the law, and spending the rest of their lives in some place where they could live freely and happily. If it is really possible for *X* to devise a plan that is so specific and life-encompassing for *Y* and *Z*, and if *Y* and *Z* are capable of executing that life-plan, in a totally coordinated and mutually supportive way, that is, in the way *X* envisaged it, then it seems legitimate to say that *Y* and *Z* would be psychologically continuous with *X*. In addition, their psychological continuity would be constituted and supported by *X*'s agency. If this far-fetched case has any plausibility, then we might say that it provides grounds for thinking that *Y* and *Z* would be the same agent, albeit spatially distributed. However, whether this idea can be defended from the perspective of the agency-based view of personal identity, and what are its possible normative and other implications, I must leave open for future discussions.

6. Conclusion

In this article, the goal was to provide an opinionated overview of the psychologically based account of personal identity and the role of agency within such an account. I followed Parfit's (1984) exposition of the psychological criterion of personal identity. Furthermore, I indicated in what way the endorsement of the psychological criterion commits one to the Reductionist View of personal identity. However, I also argued that endorsing

this view does not commit us necessarily to what Parfit calls the Extreme Claim. In this respect, I showed how the agency-based view might be useful in answering the problems posed by the Reductionist View. By relying on Korsgaard's (1989) and Bratman's (2007) views of agency, I examined the possibilities of extending the psychological criterion of personal identity with considerations related to agency, in order to see whether, and in what way, agency could vindicate practical concerns traditionally related to personal identity. I argued that, though agency-based view is promising in accommodating some of the practical concerns we relate to personal identity, it probably leaves out some of the intuitive practical concerns we might also have. I finish by sketching an example, which pertains to show in what way an agency-based view might ground personal identity that is, not only temporally, but also spatially extended.

Acknowledgments

REFERENCES

Bělohrad, R. (2014a). "Can We Do Without a Metaphysical Theory of Personal Identity in Practice?" *Prolegomena* 13: 315–334.

Bělohrad, R. (2014b). "On Schechtman's Person Life View." *Ethical Perspectives* 21: 565-579.

Baker, L. R. (2011). "First Personal Aspects of Agency." *Metaphilosophy,* 42: 1-16.

Baker, L. R. (2000). *Persons and Bodies: A Constitution View.* Cambridge University Press.

Beck, S. (2015). "The Extreme Claim, Psychological Continuity and the Person Life View." *South African Journal of Philosophy* 34: 314–322.

Bratman, M. (1999). *Intention, Plans, and Practical Reason.* Stanford: CSLI Publications.

Bratman, M. (2007). *Structures of Agency.* Oxford University Press.

Christensen, W. (2007). "The Evolutionary Origins of Volition." In D. Ross, D. Spurrett, H. Kincaid, & G. L. Stephens (eds.) *Distributed Cognition and the Will,* 255-287. MIT Press, A Bradford Book.

Davidson, D. (1967). "Actions, Reasons, and Causes." *The Journal of Philosophy* 60: 685-700.

Davis, W. A. (2010). "The Causal Theory of Action." In T. O'Connor & C. Sandis (eds.) *A Companion to the Philosophy of Action.* Oxford: Wiley-Blackwell: 32-39.

DeGrazia, D. (2005). *Human Identity and Bioethics.* Cambridge University Press.

Johnston, M. (1987). "Human Beings." *Journal of Philosophy* 84: 59–83.

Korsgaard, C. (1989). "Personal Identity and the Unity of Agency: A Kantian Response to Parfit." *Philosophy and Public Affairs* 18: 101-132.

Locke, J. (1690/1998). *An Essay Concerning Human Understanding.* London, New York: Penguin Classics.

Maslin, K. T. (2001). *An Introduction to the Philosophy of Mind.* Cambridge: Polity Press.

McMahan, J. (2002). *The Ethics of Killing: Problems at the Margins of Life.* Oxford University Press.

Olson, E. T. (2002, August 2016). "Personal Identity." *The Stanford Encyclopedia of Philosophy.* (ed.) E. N. Zalta, Stanford, California, USA.

Olson, E. T. (1997). *The Human Animal: Personal Identity Without Psychology.* Oxford University Press.

Parfit, D. (1984). *Reasons and Persons.* Oxford University Press.

Parfit, D. (1995). "The Unimportance of Identity." In H. Harris (ed.) *Identity.* Oxford University Press: 13-45.

Roache, R. (2006). "A Defense of Quasi-Memory." *Philosophy* 81: 323-355.

Schechtman, M. (2008). "Diversity in Unity: Practical Unity and Personal Boundaries." *Synthese* 162: 405–423.

Schechtman, M. (2005). "Experience, Agency, and Personal Identity." *Social Philosophy & Policy* 22: 1-24.

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

Schlosser, M. E. (2011). "Agency, Ownership, and the Standard Theory." In J. H. Aguilar, A. A. Buckareff & K. Frankish (eds.) *New Waves in Philosophy of Action*: 13-31. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

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

Shoemaker, D. W. (2016). "The Stony Metaphysical Heart of Animalism." In S. Blatti & P. Snowdon (eds.) *Essays on Animalism: Persons, Animals, and Identity*: 303-328. Oxford University Press.

Shoemaker, D. W. (1996). "Theoretical Persons and Practical Agents." *Philosophy & Public Affairs* 25: 318-332.

Shoemaker, S. (1984). "Personal Identity: A Materialist's Account." In S. Shoemaker & R. Swinburne (eds.) *Personal Identity.* Oxford: Blackwell.

Snowdon, P. F. (2014). *Persons, Animals, Ourselves.* Oxford University Press.

Swinburne, R. (1984). "Personal Identity: The Dualist Theory." In S. Shoemaker & R. Swinburne (eds.) *Personal Identity*: 1-66. Oxford: Blackwell.

Thalos, M. (2007). "The Sources of Behavior: Toward a Naturalistic, Control Account of Agency." In D. Ross, D. Spurrett, H. Kincaid & G. L. Stephens (eds.) *Distributed Cognition and the Will*. MIT Press, A Bradford Book: 123-167.

Williams, B. (1970). "The Self and the Future." *Philosophical Review* 79: 161–180.

Part V

NONEXISTENT SELF

13. On Never Been Born

MARIN BIONDIĆ

1. Introduction

In recent, well elaborated discussions about value of death, there are some questions closely related which are not my focus here. In discussions of death, when we claim that death is bad for the person who dies, we can justify that belief by a version of the deprivation theory.[1] The deprivation theory, in its core, presupposes that life is worth living and that life is a good,[2] and everything that deprives us of something good is bad for us; so death is (or can be) bad for us. If we die at the age of thirty, and the rest of our life would be good (if we had not died), then death is bad for us. Of course, life can be very nasty, and without any worthwhile future in front of us.[3] In that case we are justified in saying that death is good for us, because death does not deprive us of something good, but the opposite: death spares us from something bad. So, if life is not good, death cannot be bad.

It seems that we are prone to thinking that average lives are good, but even if life is not as good as we think it is–because people use various psychological mechanisms to assure them that things are better than they really are (Benatar 2010)–at least some lives are good enough for us to say that death can preclude such good lives, and thus be bad. If this is so, the question arises:

Is it better or worse to start living at all?

Or:

Is existence better than nonexistence for person S?

In ordinary talk we hear that being brought into existence is a gift, and religious people will add that we should be grateful to God for it. They imply

---

[1] See for example Nagel (1970), Feldman (1992), Luper-Foy (2009).

[2] At least in some cases, but many will say that it is worth living in many cases, and few that is worth living simpliciter.

[3] Like in the case of permanent torture in inhumane conditions, without any real hope for something better.

that life is good, and (no matter what quality of life it is) it is better to be than never to exist. Some pessimistic thinking people–such as Schopenhauer, who thought that the world is a punishment, or Clarence Darrow, who thought that life is completely aimless,[4] claim that it is better not to be at all. According to them, life is some kind of mistake, a bad cosmic joke, and it is not worth starting. It is better to be in a "state" of nonexistence than exist in any form of sentient life. In these reflections we can see that people mostly do not regard such value judgments as problematic. But some philosophers have inconsistent views about such value judgments. Let me explain. In this context, if we say "x is better/worse than y," we think that it is better/worse for some person S. We do not think it is better/worse for a world or in some another way impersonally. This "betterness or worseness" is person-relative. And, the argument continues, something can be better or worse for someone S only if S exists. Epicurean philosophers will claim that this is the reason why death is not bad for the person who dies.[5] But in the case of the badness of death, we can claim that there was a person S who would live if she had not died when she did. In the case of death, we compare existence with postmortal nonexistence.

By contrast, we compare and value existence with prenatal nonexistence. Or, one step further, we can compare and value prenatal nonexistence with possible existence. For example, we can say: "It is bad/good for me that I was born, and it is worse/better than not to be born at all." Or, "It is better/worse for possible people to stay in nonexistence than to start to exist." But remember, evaluation is person-relative. Who are we talking about when there is no person S who does not exist? So, when we say that "existence is better/worse than never existing," are we (at least in some cases) talking nonsense? Who is person S if S does not exist? If evaluation is meaningful (at least in some cases), how can we decide what is better/worse for person S? What should we compare it with to get the result?

2. Main Theses and The Reference Argument

When we claim that existence is better or worse than nonexistence, first of all we should specify theses. In the introduction we saw that there is more than one comparison between existence and nonexistence. And we are here interested with two of them. There are two very similar theses:

T1: Being brought into existence is better/worse than never existing.

---

[4] See Edwards (1967) in Cahn & Klemke (2008: 115-117).
[5] Known as the "existence requirement." For detailed explication and one solution of that problem see Silverstein (1980).

T2: Not being brought into existence is better/worse than existence for a nonexistent person S.

We can now put the problem of reference indicated in the introduction as an argument, and see if T1 and T2 are tenable.

Reference argument
1. If there is no person S about whom we talk, then we cannot refer to person S.
2. If we cannot refer to person S, then we cannot ascribe value judgments to the supposed "person S."
3. There is no person S who has never existed.
4. We cannot refer to a person S who never existed.
5. So, nothing is better/worse for a person S who has never existed.

Therefore, according to the reference argument, there are no rational judgments comparing the value of existence versus nonexistence for a person "S" who never exists. One old dictum is that the luckiest people are those who had never been born. On the reference argument, this dictum is nothing but nonsense, or empty words with no meaning. There are no people to whom we can ascribe such "luck." There are no people for whom nonexistence is better than the alternative. So, obviously, T2 "Not being brought into existence is better/worse than existence for a nonexistent person S," is under attack in the reference argument. T2 is meaningless.

But what is the scope of the reference argument? Can the reference argument equally apply to T1? It seems not. While T2 is about non-actual beings who never existed, T1 is about actual persons. We can refer to actual persons. So, the reference argument is not applicable to T1: "Being brought into existence is better/worse than never existing." T1 has a referent, and therefore is meaningful.

Most prominent philosophers (Nagel 1970, Parfit 1984, McMahan 1988) accept this line of thought. Nagel says, optimistically:

All of us, I believe, are fortunate to have been born… it cannot be said that *not to be born* is a misfortune. (Nagel 1970, in Fischer 1993: 67)[6]

So, it seems that Nagel implicitly says that claims about the value of our existence versus nonexistence are meaningful, although "to be" in permanent nonexistence is not something that we should value. Parfit is more precise:

When we claim that it was good for someone that he was caused to exist, we do not imply that, if he had not been caused to exist, this would have been bad for him. And our claims apply only *to people who are or would be actual*. We make no claims about people who would *always remain merely possible*.

---

[6] My italics.

> We are not claiming that it is bad for possible people if they do not become actual. (Parfit 1984, in Benatar 2010: 122)[7]

So, according to Parfit, we can refer only to actual (or would-be actual) people, but we cannot refer to people who "always remain merely possible." That would be a nonsense. McMahan continues this line of thought:

> Never existing is not something that ever happens to actual people. *A fortiori*, there are no actual people for whom never existing can be bad. (McMahan 1988, in Fischer 1993: 241)

There seems to be a difference between never existing and no longer existing. This difference allows the claim that "death (postmortal nonexistence) is bad for the person who dies" and at the same time that "prenatal nonexistence is not bad for someone who never exists." Dead people are also nonexistent entities. Every one of us knows someone who was alive, but is now dead and does not exist any more. We usually talk about these dead people, ascribing various properties to them. One of these properties is that their death (postmortal nonexistence) is bad for them, or worse than the possible life they would have lived. Are we justified in referring to such "dead people"? Is there a relevant difference between "dead people" and "people who never exist"? A dead person S is a person S who died, who was alive, and who existed but does not exist anymore. By contrast, a "person S" who never existed is not a person S at all, has never been and will never be. Nobody knows someone who is in a "state" of never existing, because this seems impossible. Nobody waits, in the waiting room of prenatal nonexistence, for his order to exist.[8]

Therefore, we can refer to a person S only if person S is actual, exists at *t*, or if person S was actual and existed previously at *t;* and we cannot refer to a person S if S is never actual and always remains merely possible. So every one of us now-living beings, and every person who was alive, is a subject and a referent of value claims about existence and nonexistence.

---

[7] My italics

[8] For the opposite argument see Yourgrau's position below in the text.
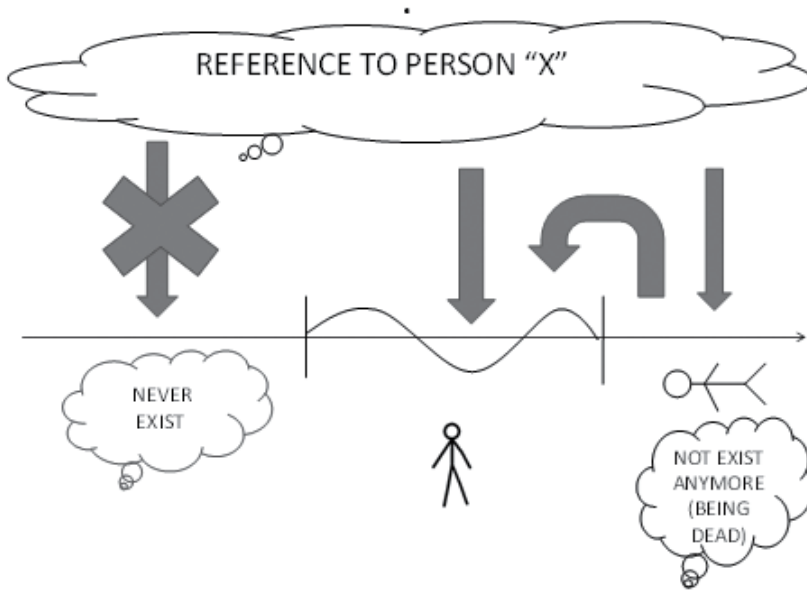
1. General schema of reference to "persons"

3. Comparative Requirement and Parfit's Solution

In everyday life we use a comparative way of thinking when we evaluate some events. For example, an anesthetic before surgery is good for us (even if we don't feel anything) because the alternative would be bad, and being in a coma is bad for us (even if we don't feel anything) because the alternative would be good.[9] This comparative way of thinking about what is better or worse for us is also common in contemporary discussions of value of death. In short, what is good or bad for us is a result of comparison of two alternatives. In the case of death we compare an actual welfare level (where a person died at some time) with a non-actual possible welfare level (where the person would live longer than in the actual world). At least in principle we can calculate the welfare level for both possible worlds and then see which of them has the higher welfare level. In other words, we can see which of the possible worlds is better. If the non-actual possible world is better (where the person did not die) then death is bad, and if the actual world is better (where person died) then death is good for that person S.[10]

---

[9] The examples are from Luper-Foy 2009.
[10] See Feldman 1991.

In short, we benefit a person S if we do what is *better* for that person S, and we harm a person S if we do what is *worse* for that person S.

But if that is correct, can we use the same comparative account in the case of evaluating prenatal nonexistence versus existence? In his analysis, Parfit considers The Full Comparative Requirement (FCR), which can be problematic for value judgment about existence versus prenatal nonexistence:

> FCR: We benefit someone only if we do what will be better for him. (Parfit 1984, in Benatar 2010: 121)

The argument goes this way:

> **The Argument From the Full Comparative Requirement** (FCR)
> 1. Person S exists.
> 2. If something is better for person S, then the alternative would be worse for person S. (FCR)
> 3. If person S had not started to exist (alternative to existence) this would not have been worse for person S (there would be no person S at all).
> 4. So, existence cannot be better than (prenatal) nonexistence for person S.

Therefore, existence is not *better* for person S because the alternative (prenatal nonexistence) would not have been *worse* for that person. If the value of existence cannot meet FCR, should we abandon the possibility of evaluating existence or coming to be existent for an actual person S? According to Parfit that is not a case:

> Causing someone to exist is a special case because the alternative would not have been worse for this person. We may admit that, for this reason, causing someone to exist cannot be *better* for this person. But it may be *good* for this person… For almost all events, if their occurence would be good for people, their non-occurence would have been worse for these people… there is one special event whose non-occurence would not have been worse for this actual person. This event, unsurprisingly, is the coming-to-be actual of this person. (Parfit 1984, in Benatar 2010: 122)

This, it seems to me, is very promising strategy. If something is not better for us, according to FCR, that does not mean that it cannot be good for us. This move from "better" to "good" allows us to evaluate the existence of an actual person S. If that were not the case, then we should remain silent about something to which we so intuitively ascribe value – our own coming into existence. Treating the evaluation of existence as a special case, "an exception to any general rule" (Parfit 1984, in Benatar 2010: 123), is completely legitimate.

When we can finally justifiably say that existence can be *good/bad* for a person S, the question remains of how to decide whether the actual existence, or coming to be, of person S is good or bad. How good/bad existence is for an existent person S does not depend on how bad/good the alternative (prenatal nonexistence) is.[11] It depends on how good or worth living life is. Parfit's view is simple and in accordance with common sense. We should sum a person S's good/bad, for example pains and pleasures[12], and then see what amount of good/bad S's life contains. This amount determines whether the life is worth living or not. If a life is good or worth living for person S, then causing person S to exist is a benefit for that person. If life in its sum is bad or not worth living, then causing person S to exist is a harm for that person. So we can accept the next claim:

T3: Being brought into existence *can* be good/bad for an actual person S.

One final remark in this context can be interesting. Do we owe gratitude to persons or other beings who are responsible for our coming-to-be? Some parents, when they are angry at their children, love to say "Show some respect (or gratitude) that I created you!" And religious people love to remind others with words "Be grateful to God, he gave you existence!" Should we? It seems not. Why? As Heyd says: "We owe gratitude to one who saves our life; we do not owe our parents such gratitude (for being saved from the limbo of nonexistence)." (Heyd 1992: 123)[13] In starting life the alternative cannot be worse: non-occurrence of life cannot be worse, nor can it be bad for person S, because there is no person S at all. "Never existing" is not something that we could ascribe to a person S.

So, it seems that value of coming to existence depends on the quality of existence for person S, and that we should not compare and value (prenatal) nonexistence for person S. But some philosophers would challenge that. Most prominent are Palle Yourgrau and David Benatar. First, I will discuss Benatar's argumentat, which is far from any optimistic picture of our existence. Then we will see Yourgrau's theory and compare both theories with Parfit's position.

---

[11] Compare the value of death. Death for a person S is bad as much as possible life would be good. Death (postmortem nonexistence) has value because of a possible-nonactual alternative, that is, existence.

[12] Let's use for simplicity hedonistic axiology.

[13] Not to exist at all cannot be bad, but being dead is something that can be bad. Of course, if we are Epicureans, then being dead is also not bad.
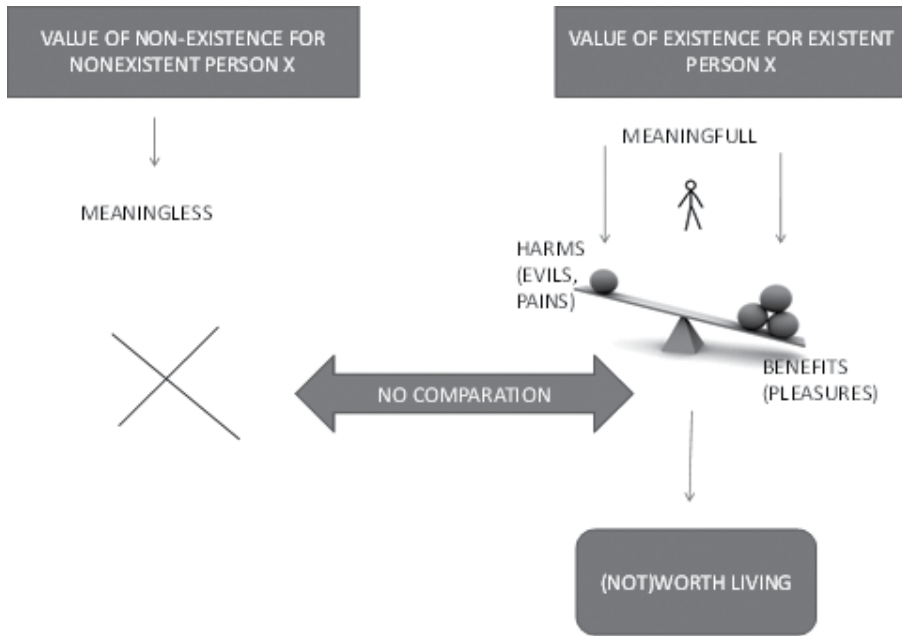
2. Parfit's account of valuing existence for person

4. Is it Really Better Never to Have Been?

It seems that most people share Nagel's view that we "… are fortunate to have been born… ." (Nagel 1970, in Fischer 1993: 67) But pessimistic people can feel that something is seriously wrong with human existence and with the existence of any sentient being. Wrong in the sense that existence is something that we should avoid. The argument for such a view is often an empirical one. The amount of various kinds of suffering is huge for most beings and significantly outweighs life's good. Of course, this is questionable. The overall result depends on many variables. Various goods and evils and their various qualities, in combination with our psychological mechanism, can make comparison very difficult. So objective assessment of the overall quality of a life, or a typical life, is highly doubtful. But what Benatar offers is independent of such a calculus. According to Benatar (1997), "it is better never to come into existence" no matter what the quality of life is. No matter what outweighs what, nonexistence is preferable to existence.

His thesis is:

T4: Being brought into existence is always a harm.

Why would someone claim this? Why would some lives, with huge amounts of pleasure and low amounts of pain and dissatisfaction, not be preferable to not existing at all? The argument is very powerful:

Benatar's argument
1. The presence of pain is bad.
2. The presence of pleasure is good.
3. The absence of pain is *good*, even if that good is not enjoyed by anyone.
4. The absence of pleasure is *not bad* unless there is somebody for whom this absence is a deprivation.[14] (Benatar 1997, in Benatar 2010: 158)

In this context, and for this purpose, premises 1 and 2 are unproblematic. We all agree with them. They are symmetrical. But 3 and 4 are asymmetrical. And what Benatar compares is 1 with 3, and 2 with 4. For everyone of us, according to this picture, in a possible-actual world where we exist, the presence of pain is bad for us and in a possible-nonactual world where we did not start to exist the absence of pain is good. In the second comparison, in a possible-actual world where we exist the presence of pleasure is good for us, but, in a possible-nonactual world where we did not start to exist the absence of pleasure is not bad. Therefore, in the first comparison the presence of pain is bad and the absence of pain is good, so we should give the advantage to nonexistence. In the second comparison the presence of pleasure is good, but the absence of pleasure is not bad, so we should not give the advantage to existence. In football jargon: on the field of nonexistence, nonexistence defeats existence 1:0. On the field of existence, existence and nonexistence draw 0:0. In the sum, nonexistence wins. So nonexistence is preferable to existence, and existence is not preferable to nonexistence. In Benatar's words:

> There are benefits both to existing and non-existing. It is good that existers enjoy their pleasures. It is also good that pains are avoided through non-existence. However, that is only part of the picture. Because there is nothing bad about never coming into existence, but there is something bad about coming into existence; all things considered non-existence is preferable. (Benatar 1997, in Benatar 2010: 162)

---

[14] My italics.

BENATAR'S COMPARATION

NOT EXISTING

COMPARATION

ABSENCE OF PAIN GOOD

COMPARATION

PRESENCE OF PAIN BAD

ABSENCE OF PLEASURE NOT BAD

COMPARATION

PRESENCE OF PLEASURE GOOD
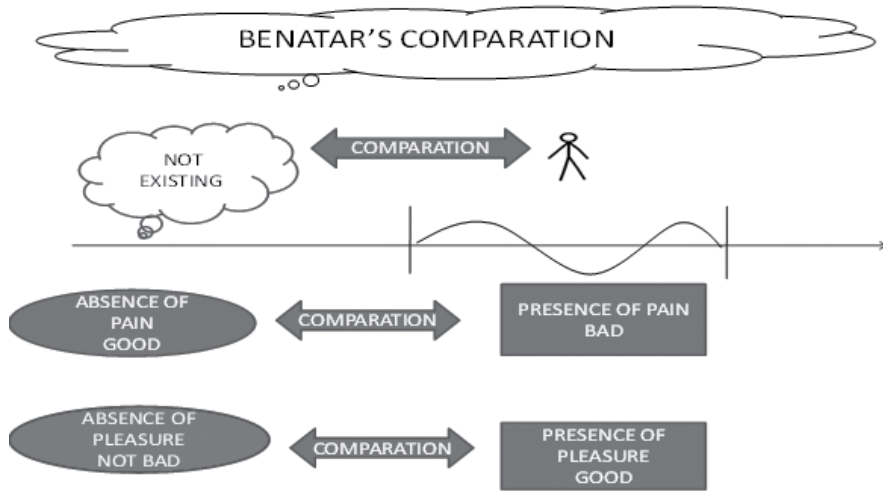
3. Benatar's comparison of (prenatal) nonexistence vs. existence

I think that the most problematic question for this argument is, for *whom* exactly is nonexistence better than existence? First of all, Benatar explicitly says that this evaluation is person-relative; it is not impersonal. Nonexistence is better than existence for actual an person S, and existence is not better than nonexistence for an actual person S. So nonexistence[15] is better than existence for every *actual* person S (for everyone of us living human beings). It would be better if we had never been. Also, nonexistence is better for every possible person S who does not actually exist but *would* exist in an alternative scenario. For every possible person S, non-existence is better than the alternative (to exist):

> … when I say that nonexistence is "better than" existence, I am not committed to saying that it is better for the non existent... that judgment is made in terms of the interests of the person *who would or has otherwise come to exist*… for any person (whether *possible or actual*) the alternative scenario of never existing is better. (Benatar 1997, in Benatar 2010: 164, 165)[16]

Is that sound? Is that plausible? When we do a comparison for person S, we compare the actual existence of person S with a possible state of affairs where person S does not exist. We have a referent for that comparison.

---

[15] Of course, it means prenatal nonexistence. Interestingly, Benatar is treating death (postmortem nonexistence) as something bad, which may seem contradictory "…there is a serious intrinsic tragedy in any death. That we are born destined to die is a serious harm." (Benatar 1997, in Benatar 2010: 164).

[16] My italics.

Every human who now lives is a subject of evaluation regarding his existence versus *his* nonexistence (prenatal). I have no difficulty in conceiving that. Similarly, in the case of valuing death for person S, we compare the actual level of welfare[17] of the now nonexistent person S with a possible state of affairs (and possible level of welfare) in which person S did not die when he did. Even if person S does not exist now, he existed, and we can say that he (who previously existed) is a subject of comparison. Every human who lived, and now is dead, is a subject of evaluation regarding his existence versus *his* nonexistence (postmortal). So it seems that we have a referent even in this case. I have no great difficulty in conceiving that too. In the first case, evaluation of existence versus prenatal nonexistence, person S is actual. In the second case, evaluation of existence versus postmortal nonexistence, person S was actual.

On the other hand, when we do a comparison for a *possible* "person S," what are we dealing with? Possible persons are persons who actually do not exist. And Benatar's thesis is applied also to possible persons who would exist. It is better for "them" to be nonexistent than to be in an alternative scenario – existent. Talk of possible people we can divide into two groups. One is "persons" who do not actually exist and will *never* exist. The second group consists of "persons" who do not actually exist, but would or will be actual. They do *not* exist *yet*. This is problematic. Here is one example.

Our ordinary talk is filled with possible people. When we lament the devastation of the environment after a possible nuclear war, we can ask whether it is better or not to create beings in such a devastated world. Is it better for "them"? Let us suppose that nuclear war destroys the Earth, leavings huge amounts of radioactivity, and that the people who remain should decide whetherto create other beings. At a meeting of the surviving people, where all the facts are brought out, half of them decide to create new people, and half of them definitively choose not to create new people. The first half after some short time make possible people actual and the second half never actualize possible people. It seems that to the point of meeting, and final decision, possibility to be existent is equal for all of non-existent people. Before the meeting we can talk about possible people in general; after the meeting we can talk about possible people who do not exist yet (but who will exist and for whose coming the survivors should prepare), and possible people who could exist but for the decision to prevent their existence. They would be actual, but something happened and they will stay in nonexistence. They will never exist (but they could have existed).

---

[17] Amount of acquired good through lived life of person S.

It seems to me that when Benatar talks of people "who would come to exist,"[18] he does not avoid a problem of nonexistent people and proper reference. It seems that we cannot distinguish a possible person "S" who will never exist from a possible person S who would exist. What exactly does that mean? Who is that person S who does not exist but would exist? To whom can we refer at this moment when we say "He would be"? Possible people are not, and if they are not, we cannot refer to them, and we cannot ascribe value judgments to them. It is just a way of thinking and talking. If that is true, we cannot claim that nonexistence is better than existence for any category of possible people. Except in one case. Only if we accept the reality of mere possibility, and we deny that this is just a way of thinking and talking, as Yourgrau did. Let's examine that philosophical option and its advantages and disadvantages.

5. The Reality of Possible People

When we think and talk in a way described in the example above of post-nuclear creation of people, it seems that we attribute some reality to mere possibility. We implicitly support Palle Yourgrau's (1987) theory on which existence and nonexistence are predicates of real beings. In short, there is a realm of being, and that realm is composed of existent beings and nonexistent beings. So we can say that a person S *is*, and that very same person S can be in a state of nonexistence or existence through time. For example, at *t1* Epicurus *is* and does *not exist*, and at *t2* Epicurus *is* and *exists*, and at *t3* Epicurus *is* and does *not exist* (again).[19] Persons are real, whether they are possible or actual. Let's examine some of Yourgrau's formulations:

> We should distinguish, therefore, between being something, being an *object*, and being an *existing* object. Existence is that property…which separates the living from the dead…we must distinguish the concept of objects-in-general from the concept of existent object. (Yourgrau 1987, in Fischer 1993: 142)

And here are some interesting observations:

> The dead, for example, are a set of nonexistents easier to grasp than the unborn. We can name specific dead people and we know many detailed facts about them, whereas it is difficult to find a single unborn whom we can isolate and refer to with a name… For myself, however, I find that this attitude[20] comes dangerously close to the sin of conflating ontology with epistemolo-

---

[18] Parfit (1984) also talks of "would be actual" people.
[19] Epicurus is real but does not exist, fictional characters, as Raskolnikov or Pegasus are not. And "…not only does not exist but could not come to exist" (Yourgrau 1987, in Fischer 1993, 144) Why? Because Raskolnikov and Pegasus are not. Only who *is*, can come into existence.
[20] He means on Nagel's attitude that we cannot say that not to be born is a misfortune.

gy. The most that the above considerations show is that we cannot know, or refer to, specific unborn. (Yourgrau1987, in Fischer 1993: 146)

If Yourgrau's theory is correct then we should return to the previously rejected thesis:

T2: Not being brought into existence is better/worse for a nonexistent person x.

We rejected T2 because there is no referent. But if nonexistent persons *are*, even if they do not exist, we have a referent, and the *Reference argument* fails. Premise 3, "There is no person S who has never existed." is wrong. Yourgrau claims that there are persons who will always remain in a state of nonexistence ("most people will never exist"), and that is unfortune for them ("have the bad luck not to enjoy existence"). (Yourgrau 1987, in Fischer 1993: 147) So there is no difference *in reality* between possible people who will never exist, possible people who would or will exist, and us, existent people. We are all real. Of course, the property of existence separates us living beings, from other possible people and dead people.

This is hard to accept for me. To say that "…the dead and the unborn are not a peculiar kind of abstract existent, but rather a perfectly ordinary kind of concrete object like you and me… ." (Yourgrau 1987, in Fischer 1993: 147) maybe philosophically defensive, but far from common sense. Yourgrau explicitly says that dead people and unborn people are "concrete nonexistent." They are not abstract entities. They are not a product of our thinking, but reality.

So are we talking about an "abstract existent" or a "concrete nonexistent" in the case of unborn people? Can we put into the same basket "dead people" and "unborn people"? I can conceive that a dead person, who was previously alive, still lives. And no matter how unlikely, I can conceive that the very same person could be restored by some almighty being or brought back by a coincidental reassembly of atoms. When I think of that, I think of a concrete person S through time. Can we imagine anything similar about an unborn person? Can we conceive of anything about an unborn person except that he (eventually) might come into existence? Maybe that is an epistemological problem, but it seems that it is closely related to ontology.

As Yourgrau noted, the dead have a modal and temporal dimension (they are possible nonexistent-existent-nonexistent object through time), and the unborn have only a modal dimension (they are possible nonexistent objects).[21] And, for Yourgrau, a modal dimension is sufficient for

---

[21] Because of that, for Yourgrau, death could be tragic, and not to come into existence only a misfortune.

reality. But when I think of reality I think of spatio-temporal reality. If something does not have, or never has had, a temporal dimension, can we really say that this entity is real? I think not. In a nutshell, I would say: no temporality, no reality. We have a temporal dimension; dead people had a temporal dimension, the unborn do not have and have never had a temporal dimension. So, I think that unborn and possible people are not real. If they are not real, we cannot ascribe existence/nonexistence value judgments to "them."

6. A Final Reconsideration

In this article I have presented four value theses about existence versus nonexistence for a person S, and the most prominent philosophers who defend or attack them. They are as follows:

> T1: Being brought into existence is better/worse than never existing.
> T2: Not being brought into existence is better/worse than existence for a nonexistent person S.
> T3: Being brought into existence *can* be good/bad for an actual person S.
> T4: Being brought into existence is always a harm.

The first thesis, T1: "Being brought into existence is better/worse than never existing." as Parfit argued, is confronted with the Full Comparative Requirement argument, and we cannot claim that existence is *better* than never existing for an actual person S (no matter what the quality of life is), because the alternative (nonexistence) is not *worse*. Also we cannot claim that existence is *worse* than never existing for an actual person S (no matter what the quality of life is), because the alternative (nonexistence) is not *better*. The alternative, never existing, is not something that can either be good or bad.

The second thesis, T2: "Not being brought into existence is better/worse than existence for a nonexistent person S" is confronted with the Reference Argument. If there is no person S, and there is not, then we cannot refer to that alleged person S. If we cannot refer to person S, we cannot claim that something is better or worse for that "person S."

The third thesis, T3: "Being brought into existence *can* be good/bad for an actual person S" is Parfit's thesis, and it seems the most plausible thesis in this group. It avoids the Reference Argument and the Full Comparative Requirement argument. We refer to some actual person S, and it seems that being brought into existence should not be better than never existing in order to be good (or vice versa).

The fourth thesis, T4: "Being brought into existence is always a harm," is Benatar's thesis, and refers to actual and would-be-actual people. If it

works as a thesis for actual people, then it is plausible. In that case we compare, for a person S, the actual world in which the person exists with a possible world in which that person does not exist. The problem, in my opinion, arises when we apply the thesis to a possible "would-be" person S. In that case we compare, for a possible non-existent person S, the actual world in which S does not exist with a possible world in which person S would exist. But in both worlds person S is not actual, and I do not know who S would be, if he never became actual. We cannot specify this "person S." So, one element of Benatar's comparison is highly questionable.

It seems to me that T1, T2, and T4, can be true only if we accept the reality of mere possibility, and that would be a high price to pay. To avoid the reality of merely possible people, the value of existence for a person S is most plausibly determinate if we restrict Parfit's non-comparative thesis T3: "Being brought into existence *can* be good/bad for an actual person x," only to actual people. What this implies for a person S, or for the average life of certain beings, is another, empirical question.
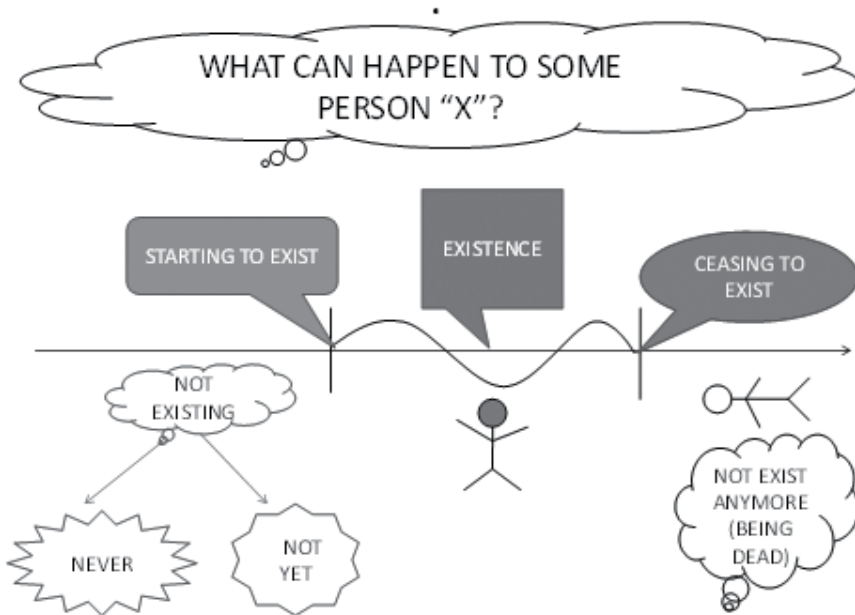
General schema – what can happen to a person

REFERENCES

Benatar, D. (1997). "Why it is Better Never to Come into Existence." In *American Philosophical Quarterly* Vol. 34, No. 3, July 1997: 345-355.

Benatar, D. (2010). "Suicide: A Qualified Defense." *The Ethics and Metaphysics of Death*: 222-245. In Benatar 2010: 307-331.

Benatar, D. (ed.) (2010). *Life, Death, and Meaning*. Maryland: Rowman and Littlefield Publishing.

Cahn, S. & Klemke, E. D. (ed.) (2008). *The Meaning of Life*. Oxford University Press.

Edwards, P. (1967). "The Meaning and Value of Life" *The Encyclopedia of Philosophy* Vol. 4. In Cahn & Klemke (eds.) 2008: 114-142.

Feldman, F. (1991). "Some Puzzles About the Evil of Death," *The Philosophical Review* 100: 205-27. In Fischer (ed.) 1993: 307-326.

Feldman, F. (1992). *Confrontations with the Reaper*. Oxford University Press.

Fischer, J. M. (ed.) (1993). *The Metaphysics of Death*. Stanford University Press.

Heyd, D. (1992). *Genetics*. University of California Press.

Luper-Foy, S. (2009). *The Philosophy of Death*. Cambridge University Press.

McMahan, J. (1988). "Death and the Value of Life." *Ethics* 99 (1): 32-61. In Fischer (ed.) 1993: 233-266.

Nagel, T. (1970). "Death." *Noûs* Vol. 4, 1: 73-80. In Fischer (ed.) 1993: 61-69.

Parfit, D. (1984). *Reasons and Persons*. Oxford: Clarendon Press.

Silverstein, H. (1980). "The Evil of Death." *Journal of Philosophy* 77 (7): 401-424. In Fischer (ed.) 1993: 95-116.

Yourgrau, P. (1987). "The Dead." *Journal of Philosophy* 86 (2): 84-101. In Fischer (ed.) 1993: 137-156.

14. Fictional Characters

IRIS VIDMAR

1. What's in a Fictional Character?

Discussions over fictional characters tend to be pushed to two extremes, giving rise to what I will call, the "puzzle of who or what fictional characters are?" On the one hand, philosophical mystery revolves around the question of non-existent objects that we can nevertheless refer to, make true or false claims about, even shed tears for. Philosophers of language, metaphysicians and logicians have spared no ink trying to explain what it is in our language that makes it possible for us to do so. How to make sense of claims such as "Emma Bovary is unhappy" given that, allegedly at least, there is no Emma Bovary, or if there is, her existence (though arguably not her unhappiness) is of a different kind than the existence of you or me? On the other hand, literary critics, fuelled by various sorts of Freudian, Marxist or Feminist theories, have done just as admirably lot of work to explain why she is unhappy, to a great extent unbothered with the fact that they are explaining emotional states of a nonexistent woman. Equally unbothered by Emma's nonexistence were millions of readers who followed her on her path to decay, some annoyed by her temper, some taking pity on her misery. How can that be?

In this paper, I set out to provide an account of the identity of fictional characters,[1] taking as my starting point the puzzling fact that although fictional characters are non-existent, we treat them as real, so much so in fact, that from our earlier days we are told stories about them. Whether we are rejoicing at the "happily ever after" that awaits the Sleeping Beauty, or are grief-stricken when little Nell dies, believing in fictional characters, taking interest in them, and, most importantly, having a more or less developed account of who they are and why they do what they do, is part of a normal, healthy cognitive and moral development.[2] It is also an indispensable part

[1] While many of the elements of my account can easily be adjusted to apply to fictional characters found in cinematic and performing art, my focus here will be limited to characters found in literary fiction. As a point of reference I will use Gustave Flaubert's masterpiece *Madam Bovary*, but note that what I have to say about it should not be taken as interpretation of this amazing literary achievement.

[2] From Darwinian explanations to psychological accounts, various theories suggest that

of literary practice[3] and literary criticism.[4] This tendency raises a host of issues, since fictional characters are neither real people (they do not exist!) nor are they like *real* people given that they have some properties (like being fictional) that people lack. This goes for those characters which are entirely the creation of a writer's imagination (think of Peter Pan), as for those which represent real world people (such as Napoleon from Tolstoy's *War and Peace*) or are to some extent based on someone real, including authors themselves.[5] How then to solve the puzzle of their identity?

I should say at the outset that many philosophers would reject the claims I made in the opening paragraph, pointing to a variety of metaphysical theories which offer different accounts of the existence of fictional characters (taken jointly, these are fictional realists). From the idea that they are abstract entities or possible people, to the idea that they are created by their authors, from the idea that they have subsistence but not existence, philosophers do not lack resources to explain how fictional characters exist.[6] Naturally, they do not exist as "real people" (even when there were real people who served, willingly, knowingly or not, as models for fictional characters) or as natural kinds such as trees or buildings, but they do exist as "creations" or "inventions" or "discoveries" (for those who prefer Platonism) made by real people (literary authors) and in that sense exist as part of the fabric of our world. You can't take Emma out for coffee, but you can have coffee while you contemplate about things she did (even if only in Flaubert's novel, that is, in the fictional world of *Madam Bovary*)[7]

---

exposure to stories is an important factor in one's moral and psychological development. See Zunshine 2015.

[3] Jennefer Robinson writes "Understanding character is essential to understanding the great realist novels… understanding character is relevantly like understanding real people…" (2005: 126).

[4] As when a critic refers to Emma as a "simple sentimental malcontent" and claims she "is miserable and her dreams of romantic love are unfulfilled in her petty provincial life married to a humble doctor …" (from the Introduction to Wordsworth Classics).

[5] How far is one willing to stretch the notion of "someone real" depends on how one feels about the claim that all, or most, literature faithfully represents the real world, at least in the sense that no matter how imaginative the writer might seem, all of his creations are traceable to something in the real world that might get somehow modified, but never so transfigured as to lose its roots in reality. Given that I am strongly committed to this claim – which I see as motivating literary cognitivism, a view according to which literature is cognitively valuable because it is a source of knowledge – in most of this paper I will presuppose that literary fiction, and by extension, fictional characters, do in fact tell us something about the real world and people who inhabit it.

[6] For an overview of realist positions, see Thomasson (2003) and Brock (2002, 2010). See Jandrić (2016) for a criticism of Thomasson.

[7] As Amie Thomasson, whose theory I am greatly influenced by, puts the point, fictional characters "are not concrete artefacts like chairs and tables, for they are not par-

which suffices for her to be part of our world, since your thoughts about her (and the thoughts Flaubert had while composing the novel) are part of this world. While I am sympathetic to the idea that fictional characters are created, and that they exist as part of our literary practice and cultural heritage (more on this below), I find this approach insufficient to explicate all that goes into character identity.

In coming up with the explanation of how fictional characters exist, realist theories for the most part ignore the fact that fictional characters are primarily part of our *artistic* practice of storytelling. Rarely do they acknowledge the fact that fictional characters – the way they are described and the role they play within the story, and in generating aesthetic experience and pleasure – are an indispensable element of the *art of literature*.[8] These theories tend to be concerned with questions of reference and denotation, truth conditions and meaning of nonexistent objects or abstract entities, rather than with the way fictional characters come to life within the established literary practices (including, roughly, writing, reading and discussing literary works). This approach – which, for the ease of exposition I will call LMS approach, since it is the approach taken by logicians, metaphysicians and semanticists – is not satisfying from the point of view of literary aesthetics (LA), which I am presupposing here. My reasons for preferring LA approach have to do with the fact that fictional characters are, first and foremost, artistic creations, and while it is to be expected that they will raise interesting questions for philosophers across the board, in talking about their identity, we should not neglect their artistic status and the fact that they originate in literary works of art. Against this context, fictional characters are indeterminate, open to interpretations, imbued with properties we recognize as human, and also with all sorts of artistic qualities, serving a specific role within the fictional world, and outside of it, as fictional characters can be a vehicle of author's irony, satire, symbolism or mockery. On my view, these are all relevant aspects of who fictional characters are, aspects which remain out of sight of those who are only concerned with their ontological status.

If logicians, metaphysicians and semanticists are guilty of occasionally at least neglecting the artistic and aesthetic aspect of fictional characters, so too are literary critics and theoreticians who sometimes seem oblivious to the fact that fictional characters are first and foremost linguistic cre-

---

ticular material objects, and (although they are created at a certain time), they lack a spatio-temporal location. No informed reader expects to meet a fictional character, or thinks that they can be found at any place at any time." (Thomasson 2003)

[8] Notable exceptions I am much in debt to are Amie Thomasson (2003) and Peter Lamarque (2009, 2010).

ations and treat them as real people. As the philosophical wisdom teaches us, even if there was a person saliently similar or even identical to Emma Bovary (in terms of her physical appearance, mental states, her character and the things she does), she wouldn't be identical to Emma, nor could we claim that Emma really exists. However, there are some beneficial lessons about the phenomenology of our engagement with literary fiction to be gained when we explore in greater detail our tendency to treat fictional characters as real people. Our natural propensity to do so speaks volumes about the way in which literature is connected to the real world, the connection understood as one of mimesis. We do not employ different sets of skills in order to understand what fictional characters are doing as opposed to understand what our fellow human beings, and we ourselves, are doing; we even have the same set of ethical, psychological and moral set of words at our disposal for thinking, criticizing, making sense of, explaining both of these.[9] Judgments of "mimetic reliability" readers make in reference to different portrayals of characters in a work show that we bring insights from the real world into our reading of fiction – part of the reason why the value of Shakespeare or Dostoyevsky so greatly exceeds that of Zane Gray or Judith Krantz lies in the fact that characters in Shakespeare's plays or Dostoyevsky's novels are much more psychologically realistic in their mental and emotional states and we as readers recognize and respond to that. This isn't to say that all characters in all great works of literature are appropriately psychologically similar to real people, but discrepancies can be accounted for by the conventions of genre, by the lack of artistic skills or by author's intentions. I do not want to give too much space to defending literary cognitivism here, (to the idea that fiction offers valuable insights into our world and our human nature) but it is important to bear in mind that, when it comes to fictional characters, it is not the *character* that is fictional. Further, recognizing the "real world" aspects in fictional characters (i.e. why we treat them as people) explains why we have emotional reactions to them. If fictional characters are "place holders" for things that can happen to us, for the emotional and mental states we can have and experiences we can undergo, it is only natural that we take interest in them and show concern for them.[10] Their destinies can easily become our destinies.

In what follows, I will propose a "multi-layered account" of the identity of fictional characters. I will claim that due of their embeddedness in narrative art, which is itself embedded in culturally determined literary practice, fictional characters have identities which are composed of various layers:

---

[9] Hagberg (2016) insists on this point, see also his 2010 and Robinson 2005.
[10] I take the notion of a "place holder" from Ema Dadlez, personal communication.

those connected to author's activities in creating them and those involved in readers' activities in responding to them when imaginatively engaged with works in which these characters appear. Once this multy-layeredness is acknowledged, it becomes easier to explain their dual nature, namely the fact that, though they are artistic creations, we often think of them as real people. However, my account will only make sense if we presuppose that philosophers are right when they make a distinction between two perspectives we can take on fictional characters. If we focus on what is going on *in the fictional world*, our perspective is internal and we treat fictional characters as real people, focusing for the most part on their portrayed emotional, psychological states, and we connect with them on the grounds of the shared similarity between their interests, predicaments and destinies, and our own.[11] If however our interest is artistically motivated and we aim at exploring *the fictional world as a work of art*, then fictional characters will remain linguistic creations imbued with aesthetic and artistic features and our interest will be in exploring their function in the overall artistic design, achieved as it is through the way they are portrayed via language, not via real world psychological make-up. It is from this perspective that fictional characters gain their artistic, symbolic, referential and cultural significance, which is an extremely relevant aspect of their identity.

2.1. A Touch of Ontology: Creating Fictional Characters and Keeping Them Alive

To ask about the identity of an artistic object is to ask about the conditions of its creation (i.e. its coming into existence), conditions of its destruction (i.e. its disappearance), conditions of its persistence or survival (how does a character survive over time), about its modal properties (which, if any, of its features are necessary), and issues having to do with individuation, that is, with distinguishing one object from the other. Given this framework, an *ontological* account of the identity of fictional characters will have to explain:

i. What does it take to create a fictional character?
ii. How do fictional characters survive through time (regardless of what happens to them in the stories they originate with)?
iii. What does it take to destroy a fictional character?
iv. How do we distinguish between different fictional characters?
v. Which of the many features of fictional characters are necessary for their identity?

---

[11] For the "two perspectives" approach see Lamarque (2009, 2010) and Thomasson (2003).

Among various ontological accounts dealing with (i) – (v) issues, all or some of them, I find Amie Thomasson's artifactualist theory the most in line with my LA approach, as Thomasson is committed to respecting our common sense beliefs about fictional characters and the norms of doing so established by our literary practices.[12] According to her, fictional characters are created at a certain time through the mental and physical acts of an author writing a literary work of fiction. They are contingent, in the sense that, had the circumstances of Flaubert's life been different, he might not had had the time to write *Madam Bovary* and the characters of Emma, Charles, Leon and others would not exist. Most specifically, fictional characters "are abstract artifacts – relevantly similar to entities as ordinary as theories, laws, governments, and literary works, and tethered to the everyday world around us by dependencies on books, readers, and authors." (Thomasson 1999: xi)

By claiming that fictional characters are abstract, Thomasson wants to stress that they lack spatio-temporal location, which isn't to say that they are of the same status as Platonic ideas – this is why they are created, not discovered, as Platonists would argue.[13] Fictional characters are found in works of fiction, but, as discussed above, we do not expect to find them anywhere in the real world (i.e. neither on the location that the narrative in which they appear specifies nor on the location where the material copy of the book itself is). They are man-made, not natural kinds or eternal objects existing in the domain of platonic ideas. Were it not for the literary (one among many cultural) practice – the practice of storytelling, or, for those who prefer Lamarque and Olsen's institutional theory of literature, the practice of literature-reading and writing – there would not be fictional characters. This claim might seem trivial, but it decisively blocks certain anti-realist views according to which there are no fictional characters given that they are nowhere to be found[14]. To paraphrase Thomasson, were it not for the practice of storytelling, it would take something of a belief in a massive deception to explain why we believe in the existence of fictional characters.

By claiming that fictional characters depend on "books, readers, and authors," artifactualist account gives us means to answer (i) – (iv). To create a fictional character, there needs to be a work of fiction, i.e. a narrative which tells a story, that gives rise to the character. In other words, fictional characters do not exist without the creative act of a writer who, through

---

[12] See Thomasson 1999.
[13] See Gaskin (2013) for a defence of platonism with respect to literary creation.
[14] I'm paraphrasing Brock here, see his Brock (2002).

manipulation of language, i.e. selection of words, creates a character and gives it a certain shape and properties. The creation of a fictional character is thus a linguistic act, one for which the author of a work is solely responsible, though, as Lamarque and Olsen showed, these kinds of acts are possible due to the institutional practice of literature. In that sense, the author brings a character into existence.[15]

Many philosophers claim that the act of naming a character is crucial for its creation, as means are given to refer to one particular character rather than the other. The opening line of *Mrs Dalloway* is a case in point: "Mrs. Dalloway said she would buy the flowers." To engage with the story, a reader simply takes it for granted that there is someone called Mrs. Dalloway[16]. Given that not all characters have names, we should recognize other resources, besides names, that authors can employ to bring characters into existence. Consider the opening sentence of *Madam Bovary*: "We were in class when the head master came in, followed by a 'new fellow,' not wearing school uniform, and a school servant carrying a large desk" (iii). In this case, characters are created and discerned by the use of a pronoun (we), by their occupation (the head master and the school servant) and by description (the new fellow, not wearing a school uniform).[17]

Giving a name, or using a pronoun or some kind of description to create a character is a first step to creating a linguistic entity readers will recognize as (sufficiently similar to) real people. Because I am interested in characters' identity, not just in what it takes to create them, I will claim that all the descriptions involving and relating to a character x are relevant for x's identity. I will have more to say about this below, for now, it is enough to say that once an author makes a decision that a work is done, the foundations of each character are determined by what is described in the story, and the linguistic descriptions that give rise to it are unchangeable (though they give rise to variety of interpretations, that is, various answers to the

---

[15] This isn't to say that we do not need an additional, psychological story to explain what goes into the creation of literary works and characters, explanation which would include author's intentions, desires and goals. Linguistic act itself is preceded by the mental act – a decision an author makes to write a story. However, while all of these aspects are necessary for the creation of a work and fictional characters, they are not sufficient, in that unless there is a linguistic act (written or oral), no one but the author himself has access to his creations.

[16] Thomasson (1999) draws the analogy with the speech act theories of language to explain how the authorial "say so" generates something into existence. See Lamarque and Olsen (1994) and Lamarque (2010) for a discussion over speech act theories and fiction.

[17] This strategy can accommodate fictional characters such as Dr.Jekyll - Mr.Hyde, and those like the nameless Monster from Mary Shelley's *Frankenstein*.

question of who that character is). I will refer to this as the "linguistic description foundation."[18]

Turning now to (ii), the existence of a character. After their creation, fictional characters no longer depend on the linguistic acts of an author, but on the existence of the narratives in which they occur (though this does not imply that they depend on any material copy in particular) and on competent and knowledgeable readers who engage with these narratives (or, in the case of oral literature, pass them on orally). Consequently, once such readers disappear, or once the works themselves disappear, fictional characters disappear too. In that sense, our answer to (iii) is the following: fictional characters can only be destroyed in the sense that they vanish from our literary horizon due to the destruction of material copies of works in which they first appeared, making it impossible for potential readers to engage with these works. In case of oral literature, disappearance of readers who have the relevant memory and knowledge of the works would bring about the destruction of characters that appear in these stories.

Another aspect of the ontological account of fictional characters concerns their individuation: fictional characters might seem diverse, but really are not. (Lamarque 2009, 2010) After all, in a sense, fictional characters are nothing but a set of properties assembled together and united under a name, and not even the most imaginative authors out there can invent new properties; they just borrow them from what they see in the real world. At best, they can imagine an original set of properties, but properties themselves – being smart, handsome, romantic, unhappy, honest, a crook, a rascal and what have you – are not, and cannot be imagined or invented. On this view, a creation of a fictional character is more a matter of "pick and choose" than a matter of creating something. Despite the surface differences in what Edgar Allan Poe, Arthur Conan Doyle, Agatha Christie, Raymond Chandler and Sara Paretsky are doing (in writing, respectively, about C. Auguste Dupin, Sherlock Holmes, Hercule Poirot, Philip Marlowe and V. I. Warshawski), they are not really creating fictional characters, since they neither created a detective, nor any of the properties associated with these characters.

---

[18] Minor potential issues can be ignored for the time being, issues having to do with potential errors in transcript or omission of words from one copy to the other (or, in the case of oral literature, errors in retelling story from one person to the next), change of word-meaning that might significantly change a description (think of gay as adjective), variations in word connotations in different languages etc. My point is, the identity of fictional characters, being tied to the narrative in which they originate, is therefore fixed by the narrative (although, as we will see below, this grounding can be extremely loose, in which case the identity of character will be very poorly grounded).

What is the power of this argument? On my view, even if authors do not, even cannot, imagine properties which would be so original as to not be susceptible to the charge considered above, that still would not imply that they are not creating fictional characters by putting together, via linguistic means, particular, aesthetically intriguing descriptions that give readers means by which to imaginatively engage with the narrative, to follow the story it tells, and most importantly, to gain aesthetically rewarding experiences from doing so. After all, those authors who are genuinely capable of doing so, go down in history as geniuses, those who fail are quickly forgotten. What actually matters, in relation to (iv), is the kind of interest we bring to the work.[19] We might be interested in assessing how an artist describes an instantiated version of a character type that exists independently of his work, i.e. how she fills in the blank space that a certain genre requires. In that case, we will focus on linguistic means that, say, A.C. Doyle employs in order to create a detective which shares some features with other (fictional) detectives – like the feature of solving crimes, outsmarting the baddies, outwitting the opponents, getting the lady, salvaging a damsel in distress etc. – but is also unique in its own way (playing the violin and smoking opium). From this perspective, our interest is in comparing and contrasting how one work falls back on the tradition in which a certain character exists. On the other hand, we can be interested in the fictional world of the work itself, in which case we will be less concerned with characters as instantiated types. What makes Sherlock Holmes so immensely fascinating as a literary achievement is only partly determined by Conan Doyle's depiction of a detective and those seeking artistic qualities of his novels will move beyond considering Holmes' portrayal in comparison to other detectives to consider the fictional world of Conan Doyle's stories.

How then to differentiate between characters? My suggestion is that a reader is capable of individuating a certain character when she (a) successfully traces its narrative of origin, (b) has a sufficiently informed understanding of what makes that particular character – character x, distinct from other fictional characters that have features in common with character x, as well as from other characters within the same work. For example, to individuate Emma Bovary from other fictional adulteresses, one needs to trace its origin to the novel *Madam Bovary*, rather than to *Anna Karenina*. To individuate it from other characters from the novel (say Charles' first wife), one needs to have a sufficiently informed understanding of how the two women are distinct. In order to gain such understanding, readers need to carefully pick up textual clues relating to each of the character and use them to construct their own image of each of them.

---

[19] See Lamarque 2010

It is an additional question, having to do with issues of modality, whether one should also trace Emma to Flaubert rather than to Tolstoy, i.e. whether it is a necessary feature of Emma that it was created by Flaubert.[20] In a sense, asking whether it was necessary that Flaubert is the author of *Madam Bovary* is the same as asking whether it was necessary that the penicillin was discovered by Alexander Fleming – once we can enjoy the benefits of penicillin being around, does it really matter that it was Fleming who discovered it? On that analogy, once we can aesthetically enjoy *Madam Bovary*, does it really matter that it was Flaubert who is to be credited with creating it? However, things are more complicated given that we tend to ascribe authors originality, innovativeness, creativity and praise them along these lines for their creations. The aesthetic achievements of Flaubert, exhibited in *Madam Bovary,* were unique at the time when Flaubert (and no one else) wrote the novel, which is an important part of the value this novel has as a literary achievement, and Flaubert as a literary artist. A word-for-word identical novel written by someone else, at some other time, would not have the same literary qualities as *Madam Bovary*. Therefore, I am more inclined towards claiming that once it is established that Flaubert wrote *Madam Bovary*, his authorship has to be acknowledged for Emma's identity, although only for her external identity (i.e. when we are interested in a work as a piece of art and in the character of Emma Bovary as an integral part of that particular novel)[21]. For her internal identity (i.e. who she is in a fictional world), the fact that she originates in Flaubert's work is less significant, as a reader who lacks knowledge of the work and character's origin can still appreciate the novel or have an understanding of who Emma is, though this understanding, and the overall experience afforded by the work, will be impoverished.

An issue far more pressing for characters' identity concerns questions such as the following: is it necessary for Emma to fall in love with Rodolphe rather than with Homais? Would she still be the same character if she cheated on Charles with someone other than Rodolpho and Leon, or only with one of them? How relevant is her infatuation with the sentimental, romantic literature for her character? This is a slippery slope argument,

---

[20] Amie Thomasson (2003) claims that it is an essential feature of a character to be brought into existence by a particular author; for counterview see Peter Lamarque (2010). See also Greg Currie (2004).

[21] This is particularly relevant when the same character figures in narratives written by different authors, such as the character of Faust. There are also cases when a certain character is "borrowed" from one literary work and inserted into another. With such cases, I would insist that the character comes with the ontological baggage given to him by the author who originally brought him into existence.

as we can modify the story in various ways, wondering whether it is still the same story, with the same characters. More formally, the question ((v) above) is, do fictional characters have core, or essential features, and how do we determine them?

I do not think there is a straightforward answer to this question. It seems we can still "get the story" and "understand the characters" even if certain episodes are absent from the work. This intuition is supported by some practical considerations: it is impossible for a reader (as well as for an author) to bear in mind the entire text of a narrative, in the process of reading as well as afterwards. Our attention as we read is selective – we might ignore certain details in order to grasp the plot line, or we might be interested in one character rather than the other, or in the aesthetics of the prose rather than the story itself. Therefore, we necessarily miss out on details, and consequently, our grasp of the characters is always porous. We are more likely to hold on to the image of Emma as a passionate adulteress and neglect the specific dynamics of her adulterous relations (with Rodolphe, she is submissive, with Leon she is dominant). In that sense, it seems that even if some episodes were absent, we would still get the story, and have a conception of who the characters are. However, from the theoretical angle, we mustn't forget that a character originates in the narrative created by an author, in accordance with her artistic vision. Therefore, we have to presuppose that every element in the story – every episode, every description, every metaphor etc. – is indispensable to that story. Every episode, in other words, has an important function within the overall artistic design of a work.[22] With respect to fictional characters, it follows that everything described in the story, in the way in which it is described, is fundamental for the story and contributes to the identity of a character. This isn't to say that a character is identified with the narrative. Characters are grounded in the linguistic descriptions of a narrative, and every episode within the narrative contributes to their identities, but their identities also depend on the literary practice and upon the activities of those who participate in them, as I will show below. "Who any given character is," on my account, is a function of author's linguistic descriptions (which she judges to best serve her overall artistic vision) and reader's constructive reading, whereby she uses the textual information, her background knowledge of the real world literary practices to come up with an understanding of this character, imbuing it not only with human-like properties, but aesthetic, symbolic, referential, etc. ones.

---

[22] I'm following Peter Lamarque (2009) here.

2. Linguistic Description Foundation of Fictional Characters

In this part I will focus on the linguistic descriptions that give rise to characters and set foundations for their identity. They also serve as an epistemic entry point for readers, who can only gain access to characters via these descriptions (more on this below). It matters little here whether an author is describing a real person, drawing inspiration from one, or whether the characters are entirely a result of author assembling together various properties she wants this character to possess. In order to bring a character to life, an author needs to first describe it in a story, i.e. ground it in a narrative, as this is the only way in which a reader can have access to it.[23]

I take the term "description" here in a rather inclusive sense, wider than usually understood, when applied only to an account of characters' physical appearance and personality traits. "Description" in this sense extends to reports of what happens in the story to each character, reports of dialogues, episodes and scenes, as each of these ascribe certain properties to characters, properties relating to characters' physical and psychological aspect, social status, belief system and the like. Given the functionality principle mentioned above, my claim is that a character's identity depends on all of these, as it is grounded in all the details of a narrative (i.e. all the information associated with a given character). Therefore, every episode is relevant for how the reader comes to understand a character. In the next part, I will have more to say on what determines the specific details of these descriptions, for now, I will give few examples of how characters are given through narrative's descriptive resources. While it would be impossible to provide a list of all the ways in which character-descriptions can be laid out, some examples will be helpful.

Consider again the first sentence of *Madam Bovary*. The fact that the "new fellow" lacks a school uniform is a subtle yet powerful way in which Charles is depicted as a man out of place with his environment and people around him, a situation he will be in for most of his life. After the opening sentence, Flaubert has the narrative "we" tell readers more details about his physical appearance and behaviour of "the new fellow," and the reader is expected to pick up clues from the text and construct an image of Charles. The fact that he is "taller than any of" the school boys indicates that he is older than his schoolmates, yet his repeated inability to introduce himself or catch up with coursework shows how poorly prepared for school he is. The sharp contrast between his shabby attire (his short jacket is tight

---

[23] Even when real people feature in fictional stories, readers "work with" descriptions provided by the author, rather than with their conception of who the person was, although they might rely on this conception to evaluate author's creation.

around his armholes, his boots are ill-cleaned) and his ludicrous cap indicates parental disharmony and neglect. As details about his parents are narrated, we learn of his mother's domination over Charles and his father's utter disregard for them both. His isolation is reflected in the contrast between those who are retelling the story, we, and "he" – the new fellow. The artistic relevance of these opening scenes is symbolic, in that they prefigure Charles' life and his way of dealing with the world; he will always be the one out of place, ill prepared, ignorant of what is happening and constantly pushed around by those around him.

Reader's construction of who Charles is depends on her successfully picking up information available from several different perspectives via which Charles is depicted. The anonymous "we" that first introduce Charles and give readers an insight into his childhood and parental relations[24] give way to a more sympathetic perspective, as when the happiness he found in marriage to Emma[25] and professional success[26] are contrasted with how Emma sees him. As Flaubert handles over the narration to her, a different image of Charles emerges, an image of a man who "could neither swim, nor fence, nor shoot...," a man who "taught nothing, knew nothing, wished nothing." (p.26) As the discrepancy between the two spouses grows, a discrepancy to which Charles is tragically oblivious as he constantly misinterprets her behaviour, Emma starts to feel more and more annoyed by him, blaming him for her misery. "Was it not for him, the obstacle to all felicity, the cause of all misery, and, as it were, the sharp clasp of that complex strap that buckled her in on all sides?" (p.68) The reader of course knows that it is not "for Charles" that she is so unhappy; given Flaubert's masterful

---

[24] "His time at school, when he remained shut up within the high walls, alone, in the midst of companions richer than he or cleverer at their work, who laughed at his accent, who jeered at his clothes, and whose mothers came to school with cakes in their muffs? Later on, when he studied medicine, and never had his purse full enough to treat some little work-girl who would have become his mistress? Afterwards, he had lived for fourteen months with the widow, whose feet in bed were cold as icicles." (p.22)

[25] "But now he had for life this beautiful woman whom he adored. For him the universe did not extend beyond the circumference of her petticoat, and he reproached himself with not loving her. He wanted to see her again; he turned back quickly, ran up the stairs with a beating heart. Emma, in her room, was dressing; he came up on tiptoe, kissed her back, she gave a cry." (p.22)

[26] "He was well, looked well; his reputation was firmly established, the country-folk loved him because he was not proud. He petted the children, never went to the public house, and moreover, his morals inspired confidence. He was specially successful with catarrhs and chest complaints. Being much afraid of killing his patients, Charles, in fact, only prescribed sedatives, from time to time emetic, a foot-bath, or leeches. It was not that he was afraid of surgery: he bled people copiously like horses, and for the taking out of teeth he had the "devil's own wrist." (p.38)

depiction of Emma, it is clear that her selfishness and self-absorption prevent her from appreciating Charles' qualities as a husband and a father. As Joshua Landy (2010) warns us, in coming up with an image of Charles, it is important to keep in mind that Emma's perspective on him is to be taken with a grain of salt.

Another technique often employed by authors to describe fictional characters involves a direct description of their mental states. The events in *Madam Bovary* are narrated from Emma's perspective, and her internal identity (who she is in the novel) starts to take shape as we learn more and more of her desires, her hopes, dreams and fears. Consider the way Flaubert describes her yearning, blind and unspecified, but so fundamental to who she is, a yearning that will later on push her into a shopping spree (which she will misinterpret as expression of her refined taste) and bed-hopping (which she will misinterpret for a true love):

> At bottom of her heart, however, she was waiting for something to happen. Like shipwrecked sailors, she turned despairing eyes upon the solitude of her life, seeking afar off some white sail in the mists of the horizon. She did not know what this chance would be, what wind would bring it her, towards what shore it would drive her, if it would be a shallop or a threedecker, laden with anguish or full of bliss to the port-holes. But each morning, as she awoke, she hoped it would come; that day she listened to ever sound, sprang with a start, wondered that it did not come; then at sunset, always more saddened, she longed for the morrow. (p.39)

Descriptions like this serve important function not only from the internal perspectives (Emma's unhappiness and a desire for "something more" explain what pushes her into adultery), but from the external one as well. It takes a somewhat sophisticated reader to connect Emma's blind yearnings and unfulfilled desires to the tradition of Romanticism, and to see Emma as a fallen romantic hero. Because Flaubert vacillates between Romanticism and Realism, Emma, as an artistic artifact, unites both. Her yearnings for a better life, for something exotic and mystic, however unspecified and blind, remain at the core of her character, pushing her around, as she is incapable of controlling her passions. Considering herself better than and superior to everyone else, Emma embodies the Romantic hero's entitlement to love, fame and wealth. However, she also embodies some features of a realist character: she is given to us in a "close up," she is firmly set in her environment which is, unlike the environment of romantic heroes, socially dense and populated with characters that occupy Flaubert's attention to a significantly lesser degree than Emma, but still sufficiently so as to offer a glimpse into the lives of a small village in French province circa 1840-ties.[27]

---

[27] See Doering (1981) for the way romanticism and realism come together in Flaubert,

Though Emma can't identify where her yearnings come from, she is more than painfully aware of where they are taking her: her progression in space and time is progression that follows from her inner states, which Flaubert conveys in impressionistic manner: "Then the lusts of the flesh, the longing for the money, and the melancholy of passion all blended themselves into one suffering, and instead of turning her thoughts from it, she clave to it the more, urging herself to pain, and seeking everywhere occasions for it." (p.68)

Dialogues and monologues are another descriptive resource relevant for depiction of characters.[28] Emma's utter incapability to care for others is best captured in her exclamation "You bother me" (p.60), when the troubled nurse asks her for help. The rottenness of Rodolphe's character is exposed in his interior monologue. We recognize his shrewdness and as he contemplates on how to seduce Emma ("With three words of gallantry she'd adore one, I'm sure of it."), given that his current mistress is "decidedly beginning to grow fat" (p.82), and we find him blameworthy for lack of ethical concerns for others, when we read that his only concern regarding the affair is "how to get rid of [Emma] afterwards?" (p.82)

There are many indirect techniques that can contribute to characters' identity, such as juxtaposition of one character against the other, as when Homais' shrewdness is contrasted with Charles' naivety, his rationalistic nature with Emma's sensual and idealistic. Name symbolism, a technique so dear to giants such as James Joyce or Charles Dickens, figures greatly in Flaubert. It is not a coincidence that Charles' surname is Bovary, a word so strikingly similar to "bovine," or that Emma is a name so often associated with English romantic literature.[29]

I suggested above that each episode is relevant for depiction of a certain character. Consider the episode in which Charles unsuccessfully performs a clubfoot operation on a stable boy Hyppolyte. While in itself a minor character, Hyppolyte's function in the story is relevant from the internal perspective, in how the clubfoot operation illuminates Emma, Charles and Homais, three characters central to the story, and from the external one, as the mockery he is exposed to because of his physical defect reflects complacent human stupidity and shallowness that so annoyed Flaubert. The episode exposes the limits of Charles' medical competence and his lack

---

and Weinstein (2009) for an account of realism and its techniques for character-depiction.

[28] In some genres, such as plays, this is the only means available. Hemingway's *Hills like White Elephants* is a short story that consists almost entirely of conversations and it is only through what is said that a reader can access the two characters.

[29] See Porter and Gray (2002).

of self-awareness with respect to it, the intensity of Homais' ambition and the strength of Emma's lust for money. From the external perspective, the episode is relevant for the structure of the novel, as it parallels the situation in which Charles and Emma first met. While their mutual mending of Mr. Rouault's leg was successful and led them into wedlock, with Hyppolyte's operation their cooperation, like their marriage, is utterly dysfunctional, causing the boy to lose his leg and Charles to lose his place in Emma's bed. Given Flaubert's family background (his father was a physician) some commentators see this episode as his commentary on the medical scene of his time and introduction of experimental sciences into medicine.[30]

Some characters may have a minor role within the fictional world, but their overall contribution to depiction of other characters might be enormous. The Blind Beggar is of marginal importance for what happens in *Madam Bovary*, but his symbolic meaning can hardly be overstated. His blindness symbolizes and reinforces the blindness of every other character: Emma is blind to Charles' goodness and devotion, to Rodolphe's deceptions, to Leon's cowardice, even to her own inability to cope with the situations she orchestrated; Charles is blind to the fact that his wife is stealing from him and is being adulterous; Homais is blind to human passions, pain and suffering; his neighbours are blind to how he instrumentalizes them; the city itself is blind to its own gullibility and mediocrity and, in a sense, people generally, Flaubert wants to say, are blind to how limited their options really are.[31] As the blindness could be an outcome of syphilis, some interpreters claim that the Beggar serves as a judgement on unrestrained sexuality and in that sense parallels Emma's feeling of being punished. Because of the way Homais, who embodies all that Flaubert finds unbearable in his social surroundings, exploits him, the character of a Blind Beggar symbolizes the helplessness of people in the face of those with financial superiority and intellectual mediocracy. Fictional characters thus often have functions that extend beyond the fictional world of a story and relate to author's aim of being ironic, satirical or didactic, or achieving aims with their works beside the artistic ones, as reflected in this critical commentary on Homais' character: "Just as there are Emmas suffering in twenty villages of France, so too are there Homaises triumphant in every city, town, and village. Flaubert bequeaths to the reader a dark vision of the future: the inevitable rise to power of the Homaises of the world, the triumph of *betise*, and the rise of totalitarian state." (p.92) On the view proposed here, this kind of functional role is another relevant aspect of who a character is.

---

[30] Porter and Gray (2002).

[31] For the relevance of Flaubert's pessimism in relation to *Madam Bovary*, see Porter and Gray (2002).

3. Fictional Characters, Linguistic Descriptions and Literary Practices

In the previous part I showed some descriptive resources available to authors for creation and description of characters; naming, direct description, description via perspective of another character, character's expression of thoughts (access to the mental states), dialogues and monologues, juxta positioning of characters, intratextual and intertextual references and name-symbolism and the function that a character has internally and externally. This list is not meant to be conclusive but illustrative, with some of the resources relating to the identity that a character has within the story and some with their aesthetic character external to the work. In what follows, I will focus on some factors, entangled and mutually dependant, that determine the choice of linguistic descriptions: those related to mimetic aspect of a work and those related to art-historic context of creation.

Mimetic dimension of literature should be understood as literature's intimate and inseparable connection with the real world: in it, we find our real world practices, institutions and cultural ways reflected, as well as our emotional, sexual, behavioural and the like patterns of human interactions. To put it simply, literature is concerned with the real world, and the real world is reflected in literary works.[32] Unique as Flaubert's heroine might seem in her futile struggles to overcome her boredom and find excitement, Emma is not unlike many of Flaubert's female contemporaries, for whom loveless, passionless marriages were the only alternatives to choices available at the time – a life of servitude, religion or prostitution. In a world where a woman could not divorce, "vote, move, open a bank account, hold a passport, or start a business without their husband's permission…" (Porter and Gray 2002: xiv), Emma's aspirations for freedom and the sense of entrapment are easier to understand. The tragedy is not only hers, as she represents a whole generation of females suffering in "twenty villages of France."[33]

How exactly mimetic dimension of a work is spelled out artistically depends on the art-historic period within which a work is created. Each art-historic period is specific in making some, but not other, artistic means available. Writing at the intersection of two periods, Flaubert could use the

---

[32] See Gibson (2007). Because of its mimetic aspect, it is often claimed that literature is a source of knowledge about the real world. I am happy to accept that claim, but it is not necessary for my discussion of fictional characters.

[33] See Porter and Gray (2002), who provide an excellent background to the social context within which Flaubert wrote Madam Bovary, and a critical discussion of his merging together the tendencies of literary realism to describe the real world and his aesthetic theory at the center of which is the form of a work.

resources of Realism – the factographic aesthetics and empirical precision of observing and describing – to convey an image of a life in a small town and the emotional commotions of his overly sensitive romantic heroine. Just couple of decades before, his Romanticist colleagues had other resources to choose from (consider the elements of the gothic novel and nationalistic tendencies in writers pertaining to this period) and as the century progressed and Realism gave way to Naturalism, forces darker than sentimental romantic literature, so detrimental to poor Emma, pushed Thérèse Raquin into the arms of Laurent LeClaire. As the public norms of what was acceptable as a topic of literature kept lowering, the way was open for authors such as Octave Mirbeau to unravel the most hidden and deviant aspects of human psychology (and only indirectly, of society), as characters such as Celestine found themselves at the mercy of sexual perverts, voyeurs and upper class gentlemen for whom extramarital sexual relations were a daily routine.

An important element of art-historic period includes genre, since conventions of genre dominant at any given point greatly influence the choice and depiction of characters, and consequently, one aspect of their identity.[34] A certain degree of formulaic consistency at the level of story creates a blank space for a particular type of a character. Consider a detective novel, which, from its birth under the genius pen of Edgar Allan Poe, centres on the character of a detective: an eccentric weirdo whose high efficiency in solving crimes is only matched by his high inefficiency in finding his way around the mores of social norms. Other such formulaic blind spots include the character of a mad scientist (gothic genre and science fiction), the "greater than life hero" (epic myths, tales of frontiers in American literature and Australian national literature), the young woman who has to guard her virtues (early 19th century sentimental novel), prince and princess (fairy tales), the pair of lovers (romances) etc. This isn't to say that authors do not experiment, break the rules and impose new directions – after all, paradigm shifts occur as much in art as they do in sciences – but certain properties of artworks, such as the choice and depiction of fictional characters in literary works, are best understood if the context of creation, and the genre conventions, are taken into consideration.

4. Fictional Characters as Representatives of Types or Classes

It is a common tendency in literary criticism, as reflected in the quote above, to claim that fictional characters represent types or classes of people,

---

[34] By claiming this however I do not want to make genre exclusively a historical category.

where "type" can designate any sort of psychological, emotional, moral, sexual or the like etiquette that can be applied to people, and "class" can be taken in its sociological, educational, geographical, economical, religious etc. meaning. Many of Flaubert's characters represent real people in this sense: Homais represents a man desperately trying to rise above his social status, on the quest for authority and power, Rodolphe represents a rich womanizer who takes advantages of his gender (something that a contemporary reader might be blind to) and social status, unbothered by the consequences of his actions and indifferent to the emotions of his fellow citizens. Charles represents naïve and timid people who lack the imagination and courage to look at reality and are therefore easily pushed around and manipulated by others. The question to consider is, if each, or the majority of fictional characters, represent some type or class of people, what does that tell us about their identities?

Consider first one difficulty. All characters are created by their authors putting together some set of features; should it happen that there are real people who can be described as having sufficiently similar set of features (yet without the feature of being fictional), it might be claimed that they are represented by those fictional characters who, in addition to being fictional, possess those same features as people in question. Yet, not only would it be incredibly unlikely to find real people and fictional characters with exactly the same set of features (with the exception of being real vs. being fictional), it would be equally hard to come up with a list of features that would completely exhaust all that goes into a real person, and all that a fictional character stands for. Are we to focus on Emma being unhappily married, and claim that she represents all unhappily married women, or should we specify this further and claim that she represents all unhappily married women who have lovers and pile up debts? In other words, how are we to identify the relevant set of features (both, with reference to fictional characters and with reference to real people) that would justify the claim that a distinctive fictional character represents a distinctive group of people? To generalize this line of thought, it can be claimed that fictional characters are too much entangled with the details of a narrative to be of interest to us as representatives of real people – any attempt to break them down to some features that would serve as criteria on whom exactly they represent fails. Some philosophers see this as a reason to reject not only claims regarding similarities between fictional characters and real people, but also claims regarding literature's ability to tell us something about the real world and its inhabitants.[35]

---

[35] See Lamarque and Olsen 1994

I think more beneficial lessons are to be gained if we consider what this difficulty tells us about the activities that go into writing and reading literary fiction. First, it reinforces our claim that authors create, rather than discover, fictional characters by putting together a certain set of features. In doing so, they are guided by their artistic vision, and the kinds of characters they create serve that vision best. Very often, illuminating some aspects of our world, and types of people, via their works, is what authors want to do. The set of features they ascribe to their characters is therefore determined by their aim of telling us something about the real people. In one important aspect therefore, it is plausible to see fictional characters as type or class representatives – this only adds fuel to the mimetic aspect of a work and further inspires the intuition that literature is cognitively valuable: if fictional characters represent real world people, we can learn something about them by engaging with fictional characters. Second, recognizing some kind of representational links between fictional characters and real people explains why there is nothing mysterious in our ability to recognize real people in fictional characters, as these characters simply hold a mirror to real people.[36]

The fact that the correspondence between fictional characters' features and real people represented by those characters is not perfect should not be an obstacle to fictional characters representing types of real people. After all, when real individuals serve as representatives of some real world type or class (for statistical purposes for example), we do not demand that they are exactly alike all the people they represent. However, it is important to keep in mind that characters' representative functions are only one of their aspects, which shouldn't overcloud the relevance they have as artistic creations, or the uniqueness they have as inhabitants of fictional worlds. The fact that we recognize some fictional characters as representing some type or class of people should be taken as one among many different layers that contributes to who they are and how they are depicted. It is important to keep in mind that the parallels between fictional characters and real people

---

[36] As a case in point, consider a critical commentary on William Dean Howells's novel *A Hazard of New Fortunes*: "Howells paints a panoramic portrait of urban life. His novel abounds in richly detailed descriptions of people representing the socio-economic spectrum, including recent immigrants, transplanted Southerners, old money and the newly rich, artists and writers. The points of view expressed by these characters include a property-is-theft socialism, a conservative Gospel of Wealth capitalism, and a remnant of the Old South's feudal aristocratic perspective. The crisis of Howell's novel, a bloody riot, reflects the harsh inequities of capitalism in the late nineteenth century and the class conflict simmering just below the surface of New York society." (Crane 2007: 161) Other interesting and illuminating studies on the role that real people have in works of narrative fiction include Head (2002), Ivanits (2008).

are, on the whole, slim, and extend only to internal perspective on a work, when we take fictional characters as real people in order to make sense of the story.[37] Readers' reactions to fictional characters extend beyond acknowledging their "real world" properties; as mentioned above, fictional characters are imbued with artistic qualities that can only be recognized and acknowledged if we take external perspective on a work.

5. Readers' Role in the Construction of Fictional Characters

I claimed above that linguistic descriptions – vehicles, as it were, via which an author creates his characters and provides information about them – are epistemic entry points for readers. Readers' task is to pick up information and text clues, associate them with each specific character and merge them together, in order to come up with an understanding of who each character is and what role it has in the fictional world, and outside of it, given its artistic properties. As textual clues are always inconclusive, undetermined and susceptible to multiple interpretations, the identity of a fictional character – who that character is – will be a matter of constructive, reflective reading, not simply a matter of author's descriptions.

It is a separate issue how these two forces work together and what is the authority of each. Some aestheticians argue that the authority of an author is absolute, in that he determines what a reader is to imagine – in other words, things are the way an author wants them to be. If this were so, the identity of fictional characters would be exhausted by the creative act of the author (though activities of readers would still be necessary for their survival, as explained by Thomasson's account). However, many aestheticians are willing to loosen up the authority of authors, some, like Barthes, even to the point of denying it completely. Derek Attridge (2015) claimed that an author creates only a text, and it is the reader, i.e. an act of reading, that realizes a given text into a literary work. If this is a proper way to think about the ontology of literature and phenomenology of reading, then we should conclude that the identity of fictional characters is more conclusively determined by the activities of readers, with author's descriptions being minimally authoritative or only causally relevant, in creating a set of sentences that, when read, give rise to the reader's construction of fictional characters. Though there are some counterintuitive consequences of

---

[37] Because we see fictional characters as real people, we can make sense of those stories which feature animal characters, and stories which feature characters who are in some salient way different from ordinary folks, such as stories in the genre of science fiction. Characters that embody abstract notions, such as the character of Death, can be understood along these lines.

this view – a potential infinity of works being one and the infinitely many Emmas, in some respect incommensurate to one another being the other – Attridge has a point in claiming that an active, constructive engagement on the part of the reader adds up to the creation of a work. On my account, the identity of fictional characters has to include both of these aspects, i.e. the fact that authors determine their features by describing them in a certain way, and the fact that readers shape characters they read about according to their own ideas, expectations, experiences, knowledge etc.

As explained above, linguistic descriptions provide an epistemic entry point for the reader, who, following up on this description and accumulating bits and pieces of information (those expressed directly and those that are only implied) available in the narrative, comes up with his own idea of who the character is. For such a construction to take place, reader has to engage with descriptive resources provided by the work itself and built up from there, applying various character traits labels, ethical judgments, concepts regarding the real world and cultural knowledge, and various artistic and value-laden concepts, to descriptions that ground the character. These descriptions are never so detailed, as to add up to a complete image of a character. We are told a lot about Emma, but it is still indeterminate whether she is a victim of her own foolish romanticism or of a social arrangement and stagnant institutions. This is one way in which characters are indeterminate: not all possible details about them can ever be given. Another way in which characters are indeterminate has to do with the fact that linguistic descriptions in which they are grounded are (like works themselves) susceptible to interpretations: it is in this part that the active, reflective reading plays a role in constructing a character's identity.

In the process of constructing the identity of characters she reads about, reader draws extensively on her knowledge of the real world, her experience (worldly experience as well as artistic/literary experience) and her knowledge of the conventions of genre and art-historical context in which the work was created. Not all readers are equipped with this kind of information and while here it is not the place to discuss how all of these factors come together in the act of reading, it is important to note that how one comes to understand, appreciate and evaluate a story (in all of its elements, including fictional characters) will partly at least depend on one's background and literary experience. Reader familiar with descriptive resources available to a realist novelist will be better equipped to appreciate the way Flaubert uses them to bring Emma, Charles and others to life, and she will be able to spot Flaubert's influence on and distinction from later generation of naturalist writers. Familiarity with art-historic context within which a work was created matters, in that it provides resources for a more

informed reading. Knowing about the limited social options available to women around 1840ties, when the story takes place, helps us understand the situation in which Emma finds herself, as well as the options she had at her disposal. Familiarity with reading protocols demanded by literature generally and different genres specifically (like adjusting to the science fiction's breach of natural laws) matters, as well as familiarity with narrative techniques available to authors (a failure to recognize unreliable or self-deceived narrators might severely hinder one's understanding of the text and one's idea of who the characters are). None of what I just said implies that readers who lack this knowledge cannot engage with a work. They will miss out on some literary qualities of a work (its symbolic meaning for example, sources of influence, textual and intertextual references and the like) and potentially formulate some faulty assumptions, but they can still follow the story and enjoy the work from the internal perspective (what happens in the story and what the characters are doing).

In addition to the factors identified above, figuring out who the characters are and constructing them from linguistic descriptions is a process that is interest-relative and depends on how engaged with the work a reader chooses to be. Consider the differences between Rodolphe and Leon. If one is only interested in providing a summary of a story, they can be identified simply by their role: "being Emma's lover" suffices to identify them. However, there are immense differences between them, differences one can only acknowledge (and appreciate their aesthetic relevance) if one pays closer attention to the kind of characters they are. To make the transition from "Emma's lover" to a more elaborated identities Flaubert gives them, reader has to engage with descriptive resources employed to describe them. When Rodolphe first contemplates seducing Emma, sufficient resources are given to conclude that he is immensely insightful and can easily understand other people's state of mind, but it is immediately clear that he is insincere, manipulative, someone who does not respect others and treats women as means for sexual gratification. On the other hand, Leon's sensitivity, reflexivity, lack of experience and sincere affection he feels for Emma make him a somewhat more likeable character, even if we detest his weakness. As the novel progresses, Rodolphe remains fixed in his hedonistic manners while Leon transitions from a romantic dreamer to an urban upper-class. From this perspective, they are as distinct characters as they are artistic creations and the fact that they share the property of being Emma's lover is the only trait they have in common.

6. To Conclude

I offered a multi-layered account of the identity of fictional characters. Borrowing from the artifactualist ontology, I explained how fictional characters come into being, survive and vanish. Analysing ways in which literary descriptions ground characters, I explained the role of mimetic aspect of literary works and the art-historic context of creation for the creation of characters. Along the way, I tackled the question of characters being representatives of types or classes, and I explained how answering questions about identity of characters is relative to the kind of interest we have in the first place. I then turned to the perspective of a reader, claiming that the process of active, engaged, reflective reading matters for the construction of fictional characters' identities. I claimed that the reading process includes "working with" descriptive resources of a narrative in a way which enables readers to recognize human traces in characters, as well as various artistic properties (external perspective). Again, how invested into this a reader is depends on her interest, background knowledge, experience etc.

I will end by noting several potential worries for my theory. First, my insistence on mimetic aspect of literature, determining as I make it to the choice of characters and their descriptions, might strike someone as having too strong a role in how I conceive of literary works (and their constitutive elements) and the aim of literary practice. It might seem that I turn the mimetic aspect of a work into its epistemic function or aim (to instruct, reveal the truth) and I then take this as work's dominant aim, with all the artistic choices secondary and relative to it. While I am sympathetic to literary cognitivism, here I only presupposed that the real world is mirrored in literature and therefore, artistic choices concerning fictional characters are partly at least influenced by that. Nothing in my account denies the relevance of fictional characters for the aesthetic pleasure derived from literature. Character descriptions play an indispensable role in the aesthetic experience provided by the work. Therefore, even those who want to separate the epistemic dimension of a work from its overall design or value, can rest satisfied.

Second, given my account of readers' activities in the construction of a fictional character, there is an element of relativism in character's identity: my Emma is not the same as your Emma, which means there are as many Emmas as there are readers. I am not too bothered by this. On the one hand, the multitude of interpretations reinforces the idea that different readers come up with different understanding of who characters are – for these readers, Emma's identities will be radically different. Second, and more importantly, on my account, characters' "core identity" remains

fixed and unchanged via its foundations in linguistic descriptions and this textual evidence, inconclusive and susceptible to interpretations as it might be, still determines their identity.

Third, there is a pressing worry that I am conflating two notions: the *ontological* notion of identity with the *psychological* notion of a character. In other words, my insistence on reader's activities being necessary for the construction of a fictional character wrongly assumes that a character is an ontological category of equal status as identity. To address this worry, let me restate that my main motivation was to solve the puzzle of who or what fictional characters are, given the LA approach, that is, given how they come to life as part and parcel of our artistic practices. Against that background, it is hard to see how else we might discuss fictional characters. Consider again the difference between Leon and Rodolphe. An account of fictional characters' identity that would not relate to their characters, internal and external, could hardly explain how they differ. Character's identity cannot be identified with the act of its creation through the words written on the page. It necessarily includes reader's constructive contribution: readers impose character trait labels based on what they read and how they understand it, thus constructing identity of characters. Perhaps the lesson here is to contemplate the connection between identity and character on a greater scale, that of relating to people generally.

Finally, because of its multi-layeredness, it might seem that there is too much that goes into identity. Moral judgments inspired by Rodolphe's womanizing competences or Flaubert's ironic commentaries on bourgeois stupidity are phenomenologically interesting and artistically relevant, but do not play a role in Rodolphe or Homais' identities. I think the way to address this challenge is to make explicit one consequence of my view, namely the fact that fictional characters' identity cannot be explicated in any neatly compartmentalized category – when it comes to fictional characters, we lack the equivalent to DNA or fingerprint method that uniquely identifies human beings. Therefore, fictional characters' identity is stretched-out on the continuum between two main points: their creation in the narrative, when they are first brought into existence in the act of being mentioned (via name, pronouns, occupation or more elaborated description) and the full-fledged account of particular character, which includes the properties it has internally and externally. How far one is willing to go on this continuum is a matter of individual choice and preferences. Given my commitment to LA approach, I left behind LMS philosophers' focus on bringing the characters into existence and their ongoing polemics over their ontological status, and I tried to show all that goes into fictional characters' identity given their place in our artistic practices.

REFERENCES

Attridge, D. (2015). *The Work of Literature*. Oxford University Press.

Brock, S. (2002). "Fictionalism about Fictional Characters." *Nous* 36, 1: 1-21.

Brock, S. (2010). "The Creationist Fiction: The Case against Creationism about Fictional Characters." *The Philosophical Review* 119, 3: 337-364.

Crane, G. (2007). *The Cambridge Introduction to the Nineteenth-Century American Novel*. Cambridge University Press.

Currie, G. (2004). "Characters and Contingency." In Currie *Arts and Minds*. Oxford University Press.

Doering, B. (1981). "Madam Bovary and Flaubert's Romanticism." *College Literature* 8, 1: 1-11.

Gaskin, R. (2013). *Language, Truth and Literature: A Defence of Literary Humanism*. Oxford University Press.

Gibson, J. (2007). *Fiction and the Weave of Life*. Oxford University Press.

Hagberg, G. (2010). "Self-Defining Reading: Literature and the Constitution of Personhood." In Hagberg, G. and Jost, W. (eds.) *A Companion to the Philosophy of Literature.* Blacwell Companions to Philosophy.

Hagberg, G. (2016). "Character." In Carroll, N. and Gibson, J. (eds.) *The Routledge Companion to Philosophy of Literature*. Routledge.

Head, D. (2002). *The Cambridge Introduction to Modern British Fiction*. Cambridge University Press.

Ivanits, L. (2008). *Dostoyevsky and the Russian People*. Cambridge University Press.

Jandrić, A. (2016). "Fikcionalni entiteti kao artefakti." *Theoria* 59, 2: 5-16.

Lamarque, P. (2009). *The Philosophy of Literature*. Oxford University Press.

Lamarque, P. (2010). *Work and Object*. Oxford University Press.

Lamarque, P. & Olsen, S. H. (1994). *Truth, Fiction and* Literature: *A Philosophical Perspective*. Clarendon Press.

Landy, J. (2010). "Passion, Counter-Passion, Catharsis: Flaubert (and Beckett) on Feeling Nothing." In Hagberg, G. & Jost, W. (eds.) *A Companion to the Philosophy of Literature*. Blackwell Companions to Philosophy.

Porter, L. M. & Gray, E. F. (2002). *Gustave Flaubert's Madame Bovary*. Greenwood Press

Robinson, J. (2005). *Deeper than Reason*. Oxford University Press.

Thomasson, A. (1999). *Fiction and Metaphysics*. Cambridge University Press.

Thomasson, A. (2003). "Fictional Characters and Literary Practice." *British Journal of Aesthetics* 43, 2: 138-157.

Weinsten, P. (2009). "Modernism." In Eldridge, R. (ed.) *The Oxford Handbook of Philosophy and Literature*. Oxford University Press.

Zunshine, L. (ed.) (2015). *The Oxford Handbook of Cognitive Literary Studies*. Oxford University Press.

Part VI

METAPHYSICS &
PHILOSOPHY OF
LANGUAGE

15. Haecceity Today And With Duns Scotus: Property Or Entity?

MÁRTA UJVÁRI

According to Kaplan's famous dictum "haecceitism is the doctrine that holds that it does make sense to ask, without reference to common attributes, whether this is the same individual in another possible world." (Kaplan 1975) So the main role of haecceity in contemporary metaphysics is to secure the transworld identity (TWI) of concrete individuals in non-qualitative terms. Since Selves are individuals, presumably concrete ones, they share in the accounts of (TWI). Note that here Kaplan is talking about a "doctrine" rather than a "property."

As to the function of haecceity in (TWI), the main idea is this: the alternative qualitative identification of concrete individuals through worlds is open to charges against the Leibnizian Principle of the Identity of Indiscernibles fleshing out identity in qualitative terms. One may argue, for example, that there are counterexamples to the principle presupposed by the qualitative account since there exist numerically distinct Leibniz-indiscernible individuals. So it seems more advisable to account for their numerical distinctness in terms of a primitive, nonqualitative thisness. In fact, the positing of haecceity is strongly connected to arguments about the failure of the Leibniz principle (PII). It is an open issue, however, whether all the alternative forms of the qualitative account of (TWI) is committed to (PII).[1]

Whatever haecceity's ontological status is, it is applied also to the identity-conditions of possible worlds or scenarios. Lewis, for example, says that "haecceitism is the claim that there are qualitatively indiscernible possible worlds." (Lewis 1986: 211) This is a strong claim since it goes with the de-

---

[1] Just think of the neo-Aristotelian position according to which natures, specific and individual, are the bearers of (TWI). These natures, while qualitative, cannot be dissolved into a mere sequence of properties falling under the Leibniz principle (PII). As to the major representatives of the neo-Aristotelian position in metaphysics see (Fine 1994, Gorman 2005, Oderberg 2011, Lowe 1999).

nial of the Leibnizian principle of identity in terms of property-indiscernibility and haecceitism is recommended as the vehicle of the rival, alternative view. The core of the latter is that there are numerically distinct while qualitatively identical things/worlds and their haecceity is responsible for their numerical distinctness.

The metaphysical function is now clear; the ontological specification of haecceity as a property comes with Rosenkrantz and recently with Diekemper. Rosenkrantz says that "an entity's haecceity is a relational property." (Rosenkrantz 1993: 104) Now the question arises what does it relate to what? To answer the question let us visit the general form of haecceity-attribution: when we say, for example, that Socrates has the haecceitistic (non-qualitative) relational property of "being identical with Socrates" then we apply the following scheme: H(I; a). In this ordered pair H is haecceity, I is the identity-predicate and a is an individual in Russellian style. The Russellian qualification means that it is not the individual with its associated properties in the typical Fregean way that plays the role of one member of the pair: the constituent is not the individual under a "mode of presentation," rather, it is the bare individual itself. So, in general, for any individual, haecceity is its relational property of being identical with itself. The most convenient way to refer to the individual is to use its proper name since this way of reference can satisfy the requirement of non-qualitativeness.

Now we have learnt a further, semantic feature of haecceity: its acceptance implies commitment to an anti-Fregean view of referring to individuals. Thus, while the very notion of haecceity seems to be extremely simple, it has a rich metaphysical profile since it goes, as we have just seen, with strong metaphysical commitments.

According to Diekemper, haecceity is an "exotic type of property," in particular, this property is "primitive and purely non-qualitative." (Diekemper 2009: 255-256) Before visiting some critical comments on the presumed non-qualitative status of a property, let us see how Diekemper answers his own question: "what is it for a property to be nonqualitative?" He starts with Adams' definition of "purely qualitative property" stating that a purely qualitative property is such that it can be expressed without the aid of referential devices such as proper names, proper adjectives and verbs. (Adams 1979: 7) Let me note that Adams' definition perfectly agree with the definition and use of the expression by other philosophers as well, say, D. M. Armstrong and M. J. Loux. They all point out that a qualitative property mixed with referential devices essentially occurring in the reference to that property qualifies only as an "impure" qualitative property. E.g. while "being the son of a king" is a pure relational qualitative property, "being the

son of an Anjou king" is an impure relational qualitative property.[2]

Unfortunately, Diekemper identifies such examples for impure qualitative properties as examples for nonqualitative properties. He writes: "So, living in a large city and being the son of a carpenter are qualitative properties, while living in New York and being the son of Henry are nonqualitative properties." (Diekemper 2009: 256) Now it is fairly obvious that being impure does not amount to being non-qualitative as well. One may safely say though that haecceitistic predicates like "being identical to Venus" or "being identical to Socrates" belong to a special, restricted subclass of impure predicates or properties formed with the identity predicate and the individual as a Russellian component as it is clear from the haecceity-scheme presented here. It is also clear that the identity-predicate essentially occurs in that scheme and cannot be replaced just by any other predicate. Maybe, haecceitistic predicates like "being present" or "being actual" can also fill the slot but no more candidates are around.

Further, when Adams, Armstrong and Loux focus on the pure/impure distinction their motivation is not to yield a definition of a haecceitistic property. Rather, they share a metaphysical realist motivation to capture the formal features of those predicates that are eligible for standing for genuine universals as opposed to mere predicates that can be formed in language. Typically, open and pure predicates are the best candidates for the role of standing for genuine universals. Evidently, a haecceitist can also draw on the pure/impure distinction made by these authors and explore it to the definition of a haecceitistic property, if there is any. However, if one is up for defending the nonqualitative character of haecceitistic properties one would be inconsistent in identifying haecceitistic properties with impure qualitative properties.

Even if the definition is amended, a host of questions suggests itself: does haecceity as a property exist also uninstantiated? More precisely, the question is whether it is possible for this property to exist uninstantiated as it is the case with standard, non-exotic properties. Further, does the notion

---

[2] See (Armstrong 2004: 13) There are "impure properties, in the sense that they are properties that involve essential reference to particulars, such as the property *being descendant from Charlemagne*." See also (Armstrong 1978). Here a distinction is made between pure and impure predicates in ch.13 § IV. Also, on pp. 146-7 Armstrong writes: "Descendent's from kings is, however, a pure predicate" while "descendant from Charlemagne is an impure predicate. It applies to a certain 'open' class of particulars in virtue of certain relations which hold between them and a certain other particular, the first Holy Roman Emperor." In a similar vein, Michael Loux writes: "a property P is impure just in case there is some relation, R, and some contingent concrete particular, s, such that necessarily, for any object, x, x has P if and only if x enters into R with s and that a property, P, is pure just in case it is not impure." (Loux 1998: 128 fn. 19)

of haecceitistic properties afford a good alternative to transworld identity (TWI) captured in Leibnizian qualitative terms? The question is highly relevant since the typical haecceitist strategy for defending nonqualitative thisness consists in seeking sound counterexamples to the Leibnizian Principle of the Identity of Indiscernibles. The point is that by demonstrating, with the help of such counterexamples, the numerical distinctness of Leibniz-indiscernible individuals there opens the path for accounting for their distinctness in terms of a primitive, nonqualitative thisness. (Diekemper 2009: 260) The most familiar counterexample is Black's thought experiment with a possible world made up of two steel globes qualitatively exactly alike and the numerical distinctness of the globes cannot be accounted for in qualitative terms. This treatment of Leibniz-indiscernible individuals seems to be taken by haecceitists as one of the main merits of the haecceitistic approach.[3]

Further, we can ask how does haecceity as a metaphysical posit connect up with the individual's specific nature and its individual nature? Can it be identified with the latter? What role does it play in individuation? More specifically, does it yield the individual via capturing its uniqueness with referential devices, or, alternatively, does it afford a complete individuation? In other words, when Socrates is claimed to have the haecceitistic property of being identical with himself, is he conceived with a thick notion or with a thin notion?

Problems with the property view is most clearly spelled out by Chisholm. He remarks that haecceity as a property is conceivable only if the corresponding individual exists. But it is clearly unacceptable that we can form the notion of a property only with the aid of the existence of a contingent being. As Chisholm says, "no property is such that it can be conceived only by reference to a contingent thing." (Chisholm 1981: 22) We can add that the case is even more serious once conceivability is turned into the basis of an existential claim; i.e. claiming that the very existence of an abstract entity, like a property, depends on the existence and the mental capacity of some contingent being. Platonists would clearly deny this.

Actualists about possible worlds and individuals are Platonic about the so-called individual concepts while not treating haecceity as a genuine property. For an actualist only actually existing individuals have their haecceity. A Russellian actualist like Adams denies that there are singular propositions about nonactual individuals. Instead of merely possible individuals actualists take their proxies, i.e. their individual concepts that

---

[3] It is also implied in Diekemper's account that the qualitative view of individuals presumably fleshed out in terms of bundles of qualitative features is vulnerable to the Leibnizian PII.

can remain uninstantiated in those worlds where the individual fails to exist. Actualists need this position in order to account for the *de re* essential features of concrete things while also accounting for their contingent existence: say, Socrates is essentially human according to his individual concept but he exists only in some worlds. The Platonic feature of existing uninstantiated can be had only by the individual concepts while haecceity always goes with an actually existing bearer. Thus actualism is committed to instantiated haecceity: for example, Adams refuses the very possibility of non-instantiated haecceity. (Adams 1981) Genuine properties, however, can exist noninstantiated as well.

Let us visit now the connection of haecceity to specific nature and to individual nature. Using the "thisness'-locution one can select an instantiation of a type: say, I want to buy *this* car rather than the other one. So the type or the specific nature is already involved; what about the connection of haecceity to the individual nature? Rosenkrantz, for example, says: "Although an entity's haecceity is a relational property, an entity's intrinsic nature includes its haecceity." (Rosenkrantz 1993: 104) It is fairly clear that the two claims in this passage are in conflict with each other: either haecceity is part of the intrinsic nature of a thing or, it is a relational property in the fashion of the haecceity-scheme Rosenkrantz is committed to. Moreover, with the first option the question arises how a purely non-qualitative item can be in tight bond with a qualitative item like the intrinsic nature: what makes the former connected to the latter? How can one guarantee that the haecceitistic property of "being identical with Socrates" is compatible only with the intrinsic nature of Socrates? Can the haecceities belonging to different intrinsic natures be swapped since no logical or conceptual connection rules out such move?

The connection between haecceity and individual nature becomes simplified due to the loose terminology of treating them as quasi-synonymous. At one place Chisholm talks about "haecceity or individual essence" with a permissive "or"-connection. (Chisholm 1976: 35) Also, Leibniz, in his *Discourse on Metaphysics* 8§ uses the same permissive locution though later he never explores the consequences of this use.[4]

The association of haecceity to individual nature may become clearer if one visits the originator's use of the term. Duns Scotus meant haecceity to be the principle of individuation. In fact, he suggested individuation

---

[4] Leibniz writes in this passage: "God who sees the individual notion or "thisness" of Alexander, sees in it at the same time the basis and the reason for all the predicates that can truly be said to belong to him." Obviously, this is not haecceitas either in the modern sense or in the sense of Duns Scotus since for either trend haecceitas is not the "basis and reason" of the predicates predicable of the subject.

by haecceity as an alternative to the contemporary medieval accounts of individuation all of which he found insufficient. In particular, he refused Aquinas' Aristotelian notion of individuation in terms of the designated matter (materia designata). Here the idea is that individuals like Socrates or Plato share the specific Form of being human but they have their individual, non-sharable flesh and blood and bones, etc. In short, it is the physical makeup that individuates tokens of the same type. Since accidents cannot individuate according to Scotus, he also refused identification by other accidental features, say, the spatio-temporal location. Also, he found that the so-called "double negation," i.e. the individual's distinctness from the type and the other tokens of its type is also incapable for individuating since the required principle must be something positive.

Here is the suggestion of Duns Scotus: "I explain what I understand by individuation or numerical unity or singularity. Certainly not the indeterminate unity by which anything in a species is said to be one in number. Rather, I mean designated unity as a this … An individual is incompossible with not being a designated this by this singularity…as it is determinately this." (Scotus 1973: 588) So, for Scotus an individual is not simply a particular that is numerically distinct from particulars of the same type; it is not something with "indeterminate unity by which anything in a species is said to be one in number." The individual qua an individual is identified by its haecceity or thisness; in this vein haecceity for Scotus connects up with individuation rather than merely dividing the type into arbitrary instantiations.

The function of haecceity for Scotus is now clear; what about its ontological status? We are told by medievalists that for Scotus haecceity is an entitas positiva. (Noone 2003: 100-128) Scotus explains his position as follows: "Just as unity in common follows per se on some entity in common, so too does any unity follow per se on some entity or other. Therefore, absolute unity like the unity of an individual … follows per se on some per se entity."[5] So, the entity-view is backed by the conviction that every unity presupposes a unity-maker with an entitative status, in a fashion similar to contemporary truth-maker claims. Woosuk Park explains the entity-view in an immanentist way: within the Aristotelian substance-accident framework the only option for haecceitas is to be an entity of some sort, more like a substance, since it obviously cannot be an accident. (Park 1990: 375-397) The reason is clear: Scotus refuses individuation by an accident.

Obviously, the property-view and the entity view of haecceity are not co-tenable. But note that the views are supposed to hold against differ-

---

[5] Duns Scotus, *Ordinatio* II., quoted by Cross, R. "Medieval theories of haecceity," Stanford entry.

ent ontological frameworks: today it is the Fregean function-argument framework of first order metaphysics, with Scotus it is the Aristotelian substance-accident framework. And the feeling of discomfort with these striking differences an be mitigated by the insight that the historical and the recent versions share a common concern for the individual; this is what motivates the positing of haecceity on both sides. The historical view was visited here briefly not only to respect the true origin of haecceitas: while we have found fault with the current property-view, we have to admit also that the entity-view is far from being impeccable. After all, why to posit an entity to each metaphysical function? This is reminiscent to the problems with the orthodox truth maker theory. So, there remains the task to find the proper ontological category for hacceity once its functional roles have been identified.[6]

Scotus stressed the singularity of the individual: "it is impossible for an individual not to be a 'this,' demarcated by this singularity, and it is not the cause of singularity in general which is sought, but of this specially demarcated singularity, namely, as it is determinately 'this'." (Scotus 1973) The problem, however, is that Forms are also singular: there is just one Form of Beauty, Courage, etc. Moreover, just like the particulars "divide" the species, to speak with Boethius, species also divide the genus in the Porphyrian tree. These points are made by J. Gracia (1988). So, then, what makes the individuality of concrete individuals different from the individuality of the Forms? Gracia's view elaborated in his book is that "individuality needs to be understood primarily in terms of the primitive notion of noninstantiability." Further: "as such, (individuality) is to be distinguished from singularity even if there is no great advantage in making a distinction between particularity and individuality."[7] In fact, I am critical of this point: I think that individuality should be distinguished from particularity.

---

[6] Rosenkrantz records the following functions of *haecceity*: 1. as a primitive thisness it helps securing identity through worlds (see also Adams's account); 2. in its semantic role it turns *de re* discourse into *de dicto* eliminating thereby the problematic *de re* locutions (see Plantinga); 3. in its epistemic role discussed by Chisholm the special status of self-knowledge is explained by grasping one's own haecceity. 4. it functions as the intension of proper names (see Plantinga, Chisholm).

[7] Gracia informs us that "Boethius …was one of the first to have made an explicit distinction between particularity and individuality." (Gracia 1988: 7) We are told that for Boethius being an individual is a metaphysical feature while being a particular is a logical feature. Though Gracia gives no further clues to this vital point we might speculate how to flesh it out. I am inclined to take Boethius's remark to mean that instantiation gives us particulars as arguments to functions in first order logic while individual natures/essences/concepts as metaphysical posits give us identified individuals.

Since universals are also singular items, singularity as a feature obviously does not select only individuals. One can agree with Gracia in this respect. "Whatever is individual is singular, but not vice versa," as he remarks. But the reason why he does not see advantage in distinguishing individuality from particularity is less obvious. Maybe he presupposes that all we can say about the individuals flows from their contrast with the universals. He says that "the language of particularity is a remnant of the Platonic language of participation; i.e. of individuals taking part in the universals. In this sense 'particular' was contrasted with 'general'."

What Gracia says may apply to particulars; but this is inadequate when the issue is the individuality of genuine individuals. Gracia admits that "epistemically we approach individuality negatively, but ontologically it is something positive."[8] Now I think that such a positive ontological posit badly needs a proper epistemic account. If we just proceed in top-down realist fashion, individuals will always be characterized only by their contrast with specific natures. But this direction of concern helps in conflating individuality with particularity. While not denying the relevance of the realist concern I claim that it targets only what it is to be a particular: i.e. what it is to be a token of a type rather than a type. I suggest a further question in the metaphysical descent: what makes a particular token of a type a genuine individual rather than a mere arbitrary token? It is evident that only this question goes beyond instantiation and latches onto the individual qua an individual.

Now we can already see the limits of the instantiation-based approach to individuality represented by Gracia. He says that "individuality is the existence of a noninstantiable instance." Gracia considers for a moment a possible objection to his account to the effect that his criterion explains only that the universal is instantiated; but it does not explain that a certain designated particular is the instance. As he admits, his criterion "is responsible for making 'man' this or that man but 'no uniqueness' is guaranteed." (Gracia 1988: 172) Unfortunately, he quickly dismisses the objection by declaring that a particular has "bare existence." That is, he evades the point by making a shift from general, second order existence to singular existence: i.e. a shift from saying that there is something instantiating a feature to the bare existence of an individual whatever feature(s) it instantiates.

Obviously, second order or general existence schemes saying that there exists something with a given property will never seize the individual qua an individual; they will always yield only a particular token of the relevant

---

[8] Gracia writes: "The view presented here, then, is that individuality needs to be understood primarily in terms of the primitive notion of noninstantiability." (Gracia 1988: 7)

type. If, however, it is admitted that individuality can be properly captured only with singular existence schemes, then the difference between particularity and individuality can not be neglected. Not appreciating the difference between individuation and particularization Gracia states that there is no great advantage in distinguishing the former from the latter.

I think that the distinction he neglects is crucial to the very idea of individuality. Gracia's instantiation-based approach is only about what it is to be a particular; but it is silent about those particulars that are susceptible of being genuine individuals. And any sound theory of individuals, among other of Selves, has to account for the feature of genuine individuality. Obviously, my presupposition here is that Selves are genuine individuals; but this seems to me quite uncontroversial.

REFERENCES

Adams, R. (1979). "Primitive Thisness and Primitive Identity." *The Journal of Philosophy* 76, No. 1.

Adams, R. (1981). "Actualism and Thisness." *Synthese* Vol. 49: 3-41.

Armstrong, D. (1978). *A Theory of Universals. Universals and Scientific Realism* Volume II. Cambridge University Press.

Armstrong, D. (2004). *Truth and Truthmakers*. Cambridge University Press.

Chisholm, R. (1976). *Person and Object.* London: Allen and Unwin.

Chisholm, R. (1981). *The First Person: An Essay on Reference and Intentionality*. Minneapolis: University of Minnesota Press.

Cross, R. "Medieval theories of haecceity." Stanford entry.

Diekemper, J. (2009). "Thisness and Events." *The Journal of Philosophy,* Vol. CVI. No. 5: 255-276.

Fine, K. (1994). "Essence and Modality." *Philosophical Perspectives* 8: 1-16.

Gorman, M. (2005). "The Essential and the Accidental." *Ratio* 18: 276-289.

Gracia, J. (1988). *Individuality: An Essay in the Foundations of Metaphysics*. State University of New York Press.

Hyman, A. & Walsh, J. J. (eds.) (1973). *Philosophy of the Middle Ages*. Indianapolis: Hacket.

Kaplan, D. (1975). "How to Russell a Frege-Church." *The Journal of Philosophy* Vol. 72.

Lewis, D. (1986). *On the Plurality of Worlds*. Blackwell.

Loux, M. (1998). *Metaphysics: A Contemporary Introduction*. Routledge.

Lowe, E. J. (1999). *The Possibility of Metaphysics. Substance, Identity and Time*. Oxford: Clarendon Press.

Noone, T. (2003). "Universals and Individuation." In Williams T. (ed.) *The Cambridge Companion to Duns Scotus.* Oxford University Press: 100- 128.

Oderberg, D. S. (2011). "Essence and Properties." *Erkenntnis* Vol. 75. (1): 85-111.

Park, W. (1990). "Haecceitas and the Bare Particular." *The Review of Metaphysics* Vol. 44: 375-397.

Rosenkrantz, G. S. (1993). *Haecceity. An Ontological Essay*. Kluwer: Dordrecht.

Scotus, D. (1973). *Ordinatio* II. In Hyman, A. & Walsh J. J. (eds.) *Philosophy of the Middle Ages*. Indianapolis: Hacket.

Scotus, D. "The Oxford Commentary on the Four Books of the Sentences." In Hyman, A. & Walsh J. J. (eds.) *Philosophy of the Middle Ages*. Indianapolis: Hacket.

16. Who Am I?

ARTO MUTANEN

1. Introduction

The question "Who am I?" may arise into our minds from time after time while we, as humans, contemplate ourselves and our lives. The question particularly comes to mind when significant changes occur in our lives. Such a change need not be anything of particular kind; for example, it may occur to a child, to a teenager during puberty, or to an adult when he or she encounter a serious illness. However, what a person who is asking this question is looking for as an answer is not clear. What kind of answer is expected? What kind of answer could be understood as conclusive?

This is not a single question but a cluster-question to which different kinds of answers are expected. The answer which satisfies the asker depends on several different things. For example, the following are good examples: A small child may be asking what kind of member of society he or she is. An adult who is facing adversity may be looking for direction in life. An old man may be looking for the meaning of his whole life. All these are possible frameworks in which the question arises, the question is connected to the fundamental vulnerability of human beings.

In his book *Ecce Homo* Nietzsche had a subtitle "How One Becomes What One Is," which gives the impression that we have a kind of essence. The impression may be wrong in the case of Nietzsche, but still essentialism is not a dead approach. It takes place in philosophy and in everyday thinking. Nietzsche as a pre-existentialist was one who made Sartre-like existentialism possible: Sartre said that existence comes before essence. As humans we have dual being: As existent objects we have our own being (being-in-itself) and as self-conscious humans we have human-like being (being-for-itself). According to Sartre we humans carry out ourselves. (Sartre 1956)

To be a human is very problematic: We ask the question "Who am I?" but no answer satisfies us. This was expressed by Nietzsche in the introduction to his *On the Genealogy of Morality* as follows:

> We are unknown to ourselves, we knowers: and with good reason. We have never looked for ourselves, – so how are we ever supposed to find ourselves? (…) 'Who are we, in fact?' and afterwards, as I said, we count all twelve reverberating strokes of our experience, of our life, of our being– oh! and lose count… We remain strange to ourselves out of necessity, we do not understand ourselves, we must confusedly mistake who we are, the motto 'everyone is furthest from himself' applies to us for ever, – we are not 'knowers' when it comes to ourselves… (Nietzsche 2006: 3)

As Nietzsche shows, the fundamental question "who am I?" is not answerable. However, as humans we do not stop asking the question – it is the question that comes into our mind repeatedly. The recurring questioning refers to our deep need to have control over things. (Pihlström 2015)

The unanswerability of the main question of this paper, which is expressed in its title, makes it more attractive. We are interested in ourselves: The question "Who am I?" searches for some kind of information about myself (the questioner). However, different people are looking for different kinds of answers. Moreover, the question is not a request for the same kind of information for the same person at different ages. The question can be understood from different points of view. In the following we will give certain characterizations to the questioning, which will provide a sketch of the different kinds of meanings the question has.

2. Knowing Who

The question "Who am I?" requests an answer which allows the questioner to say truthfully "I know who I am." In philosophy, the questions that search for who someone is are special kinds of questions. They are questions about identity but not in its usual philosophical or logical sense. (Boer and Lycan 1986: xi) However, the class of question is extremely interesting to us.

Hintikka (1962) considers phrases like "knowing who" and the like. Hintikka's consideration is closely related to our main question but it relates it to a more general philosophical framework. So, even if Hintikka's approach does not give us the tools to tackle our main question directly, it is still valuable for us in the present paper. Hintikka (1962: 108) considers the specific question "When is it true to say 'a knows who b is'?"

Hintikka (1962: 108) writes: "If you ask me 'Who was the teacher of Antisthenes?' and I reply 'The teacher of Antisthenes was the same man as the teacher of Aristippus', my answer does not necessarily help you to know who the teacher of Antisthenes was if you may fail to know who the teacher of Aristippus was." This shows quite clearly what Hintikka had in mind when he was considering the "knowing who" questions: The intention is

to characterize how we can identify the person we are considering. We ask who-questions if we do not know who somebody is. These questions are seeking information that allows us to identify the person.

Hintikka characterizes the problem as one of identification: How can we identify the person adequately in each possible case? That is, the intention is to learn to distinguish the person (the individual) in different contexts. So, if I know only that "the teacher of Antisthenes was the same man as the teacher of Aristippus," it does help me to identify the person, but if I get to know that the searched person was also the teacher of Plato then I can identify him more precisely (if I know who Plato was). The latter piece of information helps me to better identify the person I am searching for in different contexts.

The individuals that are considered here are ordinary extensional individuals. In this sense the ontology remains "thin": Ontology is not over-stretched by unconventional entities. Still, we have a philosophically important problematic under consideration. Even if the individuals are usual extensional ones, they occur in intensional (or modal) contexts, which make the context in which individuals occur opaque. Contexts that are not transparent are of extreme philosophical interest. How we can refer to the entities as opaque contexts? Kripke (1972) believes that proper names do the task. However, as Hintikka's analysis of the who-questions shows, proper names are not a direct solution to the problem. The other central topic is the role of quantifiers in opaque contexts. Quine (1960: 197) says that "we can legitimize quantification into modal position by postulating that whenever each of two open sentences uniquely determines one and the same object x, the sentences are equivalent by necessity."

The condition given by Quine is very strong. Quine (1953: 142) calls these kinds of contexts referentially opaque, which is stronger than mere opacity. According to Quine, referential opacity means that we do not have ordinary objects, which entails, for example, the failure of existential generalization. Moreover, this "contravenes the very meaning of identity." (Quine 1970: 78) The Quinean solution to the problems connected to the use of language within "referentially opaque" contexts seem to be too general. In fact, we must consider more closely the character of opaque contexts that shows that opacity does not entail the "global" difficulties mentioned by Quine. The use of language is "local" and there are methods that allow us to manage the use of languages with such local situations.

The opacity of the context does not mean that there would be some kind of "intentional entity" or that the objects we are speaking about would be of any special kind. The key notion in opening up the problem is the modal profile of a sentence. (Hintikka 1969: 129) To better understand the notion,

let us consider an example given by Hintikka (1962). He considered the sentence "Watson knows that Mr. Hyde is a murderer." Watson's knowledge does not give him information which allows him to infer that the murderer is Dr. Jekyll. The reason is that

> Although "Dr. Jekyll" and "Mr. Hyde" in fact refer to one and the same man, they refer to different men in some of the "possible worlds" we have to discuss what Watson knows and does not know. (Hintikka 1962: 110)

The possible worlds in which the referents are not the same individual are neither actual nor some "odd" worlds, but worlds that characterize the knowledge Watson has. Within the worlds the murderer is not uniquely fixed, which means that the knowledge is not conclusive. But if Watson knew that Dr. Jekyll is the same person as Mr. Hyde, then the intended conclusion would follow, i.e., the names would be interchangeable. However, if Watson did not know who Dr. Jekyll is, Watson would need some further information to get to know who the murderer was. So, Watson has proper *de dicto* knowledge but not *de re* knowledge. (Hintikka: 1962) Hence, even if Watson knows that Mr. Hyde was a murderer and if he knows that Dr. Jekyll is Mr. Hyde, it does not follow that Watson would know about some actual person and that he is the murderer. This is closely connected to the problem of quantification into a modal context, which was discussed by Quine. The reason is not the referential opacity of the context, but that the information expressed does not give enough information to determine the referent in the actual world (or in the "reference world"). (Hendricks 2001, Lewis 1973)

The basic idea is that the person asking the question acquires more appropriate knowledge such that the knowledge uniquely determines the individual in question, within the relevant context. Moreover, the knowledge has to be proper knowledge such that it is true, i.e., it determines a unique individual, which is a special case of more general identification of individuals. The notion of identification refers to the method or the technique of selecting an intended individual. The question is "How to recognize an intended individual?"

Quine (1969: 8) speaks about the individuative use of language which can be learned well before we master "the ins and outs of our adult conceptual scheme of mobile enduring physical objects, identical from time to time and place to place." This shows that the identification may not be confused with identity: identification and identity are closely related notions but they operate at very different conceptual level. More clearly, identification is a methodological notion and identity an ontological notion. (Gleason 1983: 910; Quine 1969: 55)

The individuative use of language refers to a larger methodical aspect that can be seen behind the answering of who-questions. However, it may seem that the text above does not touch on the fundamental question of the paper at all. There may be several reasons for this. One reason may be the bifurcation of a person into mind and body which is customarily expressed as person's "two lives and of his two worlds by saying that the things and events which belong to the physical world, including his own body, are external, while the workings of his mind are internal." (Ryle 1949: 12). The idea that we have these two different parts is so deeply rooted in our philosophical and everyday thinking that we have to consider it a little.

3. Mind-Body Problem

A person who is asking about his or her identity is not satisfied by answers that identify him or her by factors that can be understood merely as "material" or "social." We are something more than mere biological machine-like animals. It is difficult to characterize the property which separates us from other animals. For example, Wilkes (1988: 23) lists the conditions for (human) personhood that all are subsumed into mental aspects of humanity. This kind of thinking is anchored in our understanding about humanity and about human individuals themselves. Usually this is connected to Descartes's dualism, according to which the "mind is not part of the physical world at all." (Perry, Bratman and Fisher 2010: 239) The dualistic view was also expressed by Descartes very explicitly: "I have a clear and distinct idea of myself – insofar as I am a thing that thinks and not an extended thing – and because on the other hand I have a distinct idea of a body – insofar as it is merely an extended thing, and not a thing that thinks – it is therefore certain that I am truly distinct from my body, and I can exist without it." (Descartes 1980: 93)

According to Descartes, our identity is essentially connected to our mental part. In a sense, his *cogito* argument is intending to show this. I can doubt all the physical things, but while doubting there is someone who (or something which) is doubting. This is not an empirical fact but a logical fact. So, no demon, despite how clever the demon might be, can deceive the doubter:

> But there is a deceiver (I know not who he is) powerful and sly in the highest degree, who is always purposely deceiving me. Then there is no doubt that I exist, if he deceives me. (…) Thus it must be granted that, after weighing everything carefully and sufficiently, one must come to the considered judgment that the statement 'I am, I exist' is necessarily true every time it is uttered by me or conceived in my mind. (Descartes 1980: 61)

The existence of the doubting mind is not eternal but it is (necessarily) the case only when one is doubting. Hence, Descartes did not assume that the *cogito* argument would prove the existence of an eternal and unchanging "I." The character of the "I" is to be studied. (Hintikka 1963)

Descartes was studying the fundamental character of the human being. That is, he was considering the ontology of the human being: what is the essence of being human? Hence the question is not the main question of the paper but "What then am I?" (Descartes 1980: 63) Descartes' answer "a thing that thinks" (1980: 63) is a conclusive and well-defined one. However, this is not conclusive to the questioner who is looking for his or her place in society or for the meaning of life; that is, Descartes does not give an answer to the main question of the paper.

Descartes' solution to the mind-body problem can be characterized as interactive dualism, which refers to the interactive character of the two sides of the human being. (Perry, Bratman, and Fischer 2010: 240) However, the mind is "united to the whole body, nevertheless, were a foot or an arm or any other bodily part amputated, I know that nothing would be taken away from the mind." (Descartes 1980: 97) So, the identity of a person is characterized by the mind. Moreover, some of the properties that are nowadays understood as mental were characterized by Descartes as bodily:

> Besides, I believe that this power of imagining that is in me, insofar as it differs from the power of understanding, is not a necessary element of my essence, that is, of the essence of my mind; for although I might lack this power, nonetheless I would undoubtedly remain the same person as I am now. (Descartes 1980: 90)

The dualistic Cartesian solution is not the only possibility, and today it is not the prevailing one. Regardless of what approach we have toward the mind-body problem, we have to understand that the human being incorporates these two aspects and hence also some kind of dualistic identity. For example, the question about the identity of a person in the case of schizophrenia, such as Dr. Jekyll/Mr. Hyde, or Putnam's example of brains in a vat, show the dualistic – or double – character of human beings. All these are examples that force us to carefully consider the identity problem. A thought experiment can be used to analyze these kinds of examples. (Mach 1976; Wilkes 1988)

However, the identity of a human being is not merely a momentary identity but an identity over time. How does the identity of a human individual remain from one time to another? Quine (1952; 1953; 1960; 1969) links the problem to "a popular riddle" which he connects to "the problem dated from Heraclitus, who said "You cannot step into the same river twice, for fresh waters are ever flowing in upon you." In ordinary language, the

term "same" has several different meanings. Only in special cases does it refer to identity. (Carnap 1967; Ayer 1976) Quine characterizes identity as momentary identity, similarly to Descartes, above. However, it must be explicated how these different momentary identities are related to each other, which include at least two aspects, namely the persistence question and question about the same person. (Olson 2015: 5; Russell 1918: 149)

However, there is no need for us to consider this subject in any greater detail here. It is important for us to recognize that the mind-body problem is related to an ontological problem and the question we are interested in is not (merely) ontological. According to Gleason (1983: 912) the question "Who am I?" is not an ontological one in the way Descartes understood his mind-body problem; instead it is a question about locating oneself in society. The ontological question about human beings was the real question behind the mind-body problem, which can be understood as a question about identity: What am I? The question about the location of oneself in society has also been referred to as an identity question, even if it would be more appropriate to ask about identification.

4. Identification

The question of identification is easily confused with the question of identity. That is, the identification of the questions (of identity and of identification) is not easy to do:

> Identification turns out to be one of the least well-understood concepts – almost as tricky as, though preferable to, "identity" itself; and certainly no guarantee against the conceptual difficulties which have beset the latter. (Hall 1996: 2)

It is important to characterize the identity and identification question: What kinds of questions are these? What kind information are these questions looking for?

Questions about identity look at the ontological characterization of what entity is. In his philosophy, Descartes characterizes the mind-body problem in this sense. A human being is fundamentally a mental entity. Descartes (1980: 89) also discusses the relationship between mind and body and examines "the difference between imagination and pure intellection," which shows the difference between mind and body. This shows how one can "intuit by powers of discernment" the presence of the notion. Descartes called this "imagining," which is very similar to Kant's notion of intuition. (Hintikka 1973: 117)

Identification as a question about locating oneself in society may turn "out to be one of the least well-understood concepts" because, in princi-

ple, there is no effectively characterized class of questions that fix the class identification questions and class of answers to the identification questions. However, it is possible to understand the kind of identification questions. One extremely good example is presented in the *Encyclopedia of the Social Sciences*, published in the early 1930s, which characterizes "Identification" as dealing "with fingerprinting and other techniques of criminal investigation." (Gleason 1983: 910) This shows that identification refers to methods or even techniques of determining the person. In the section "Knowing Who" above, the main topic was identification. Nietzsche (1988: 33) understood the question "Who am I?" as an identification question, which can be seen from the answer he gives to the question.

The question about identification is one about methods and techniques to define individuals. So, the philosophical discussion of identification is a methodological discussion; hence we should search for methods of identification, i.e., search for answers to the question about "knowing who." A detective asks "who is the murderer?" – instead of finding an ontological characterization of the murderer, the detective identifies him or her. The detective is attempting to distinguish the murderer; he or she is looking at the information which allows him or her to be able to truthfully say "I know who the murderer is." This is just what is meant by identification in this context: The detective wants to identify the murderer, which is necessary in order for detective to arrest the murderer.

The who-question can be understood as an identification question. However, who-questions do not constitute a single class of questions. According to Russell we have two kinds of knowledge, namely knowledge by acquaintance and knowledge by description. These are based on different types of information or, more generally, two different kinds of identification of individuals. Knowledge by acquaintance is based on direct information from the entity. Basically this is observational information. Such a method is in an obvious sense person-centered. However, the person-centeredness means that the coordinates of the knowledge are anchored to the subject of knowledge. This is the phenomenal information referred to by Russell, Carnap, and the Logical Positivists, among others. The knowledge by description is based on a "true description" of the object (Russell 1918: 54) which provides an objective characterization of the individual.

Hintikka (1969) generalizes Russell's approach. According to Hintikka, there is no single method but a wide variety of approaches to individuation, all of which are part of the different branches of philosophy, such as the philosophy of psychology, biology, and physics. (Hintikka 1969: 170) This range of methods explains why the notion of identification seems such a messy one. However, we have to distinguish between a general method-

ological approach (method of identification) and a specific methodical approach to individuate a singular entity (method of individuating), which is explained by Hintikka and Hintikka as follows:

> Instead, we would have to assume, for each primitive expression of our language, a (partial) function ('meaning function') which specifies its reference (if any) in the different scenarios ('worlds'). For instance, what was a name, with a single individual as its reference, now has as its meaning (reference) as a function ('individuating function'), which for each given world defines the embodiment of the particular individual in question in that given world. (A way of visualizing such a function is in the form of an imaginary "world line," which connects these several embodiments of the same individual with each other.) The totality of such world lines defines what counts as a method of identification for individuals. (Hintikka and Hintikka 1969: 76)

The idea is that a given method distinguishes an individual in a different context of application (possible worlds). According to Hintikka (1969) there are different kinds of methods of individuation: Perceptual methods, which correspond to Russell's knowledge by acquisition, and physical methods, which correspond to Russell's knowledge by description. The method is not a single method but rather classes of methods. In fact, the character of the methods is very multidimensional, which means that the practical application of the general approach is not easy. However, this does not reduce the theoretical interest in the conceptual framework connected to the whole approach. This allows us to classify the different approaches methodologically, especially when we recognize that several empirical studies of identity are in fact studies of identification via perceptual methods. Of course, the methods (in the sense of method of individuation) used in such studies vary from case to case.

Russell (1912: 28) refers to a different method of identification when he speaks about our identification (acquaintance) of ourselves. He explicitly mentions "the outer senses" that belong to the perspectival method but also to memory, which belongs to physical methods. Moreover, he mentions "that we have acquaintance with Self," which indicates a kind of "direct observation" of oneself. He recognizes the multiplicity of the methods, which can be seen from the characterization of the person as a constructed empirical relation "ascertained by analysis." (Russell 1918: 150) The essential thing for us is not to characterize or problematize the relation analysis as such, but to recognize that the relation is intending to do the same methodical task as the individuating functions as referred to by Hintikka. (1969: 101) So, the relation is a kind of "theoretical sum" of all the possible empirical observations. Of course, it can be critiqued that such a method does not give us the metaphysical person. Identification does not refer to search for a metaphysical person but to intention to distinguish a person in different contexts. (See also Russell 1918: 148)

Gleason (1983: 913-914) refers to the book *The Identity Society*, published in 1972, which argued that the notion of identity had "come to mean so many things that, by itself, it means nothing." The inflated use of terms makes language an empty tool and eventually it will cease "to perform the function of a verbal sign." According to Gleason, the notion of identity in social sciences is a relatively new one, and it is connected to "ethnic identity" and the problems related to it. However, Freud introduced the notion of identity "to designate the process by which the infant assimilates to itself external persons or objects." (Gleason 1983: 915) This Freudian notion is quite in balance with the notion of identification that we have already discussed.

5. Possible World Semantics

Identification is a notion which can be applied to many different contexts. Moreover, we have seen that the notion of identification gives a methodological structure to the argumentation. However, the complexity of the conceptual situation is enormous. So, there is need for further conceptual and methodological clarification.

Basically, we do not face the identification problem as a messy and holistic problem but, as the examples from the literature show, the identification problem we face in practice is a "localized" one. Gleason (1983) shows several identification problems that can be understood as local. For example, all problems in which a person is searching for his or her place in society are in this sense local: A person is not looking for some fundamental holistic understanding about the self; he or she is trying to localize him or herself in society. I am not trying to belittle the problem: it is likely to be very important for the person. Still, it is true that the individual is not looking for some kind of holistic answer, but something that helps them to feel that their life is meaningful – my membership of society is acknowledged: I know who I am. In asking these kinds of questions, the mind and body are not separated but closely connected, which shows why dualism, as introduced by Descartes, is understood as a spurious approach.

Possible world semantics is a general method of modal notions. Modal notions can be characterized as notions that refer to several different possible worlds. That is, possible worlds semantics is a "natural" semantics for modal notions. For example, b knows that p (in world w) "if p is in all possible worlds compatible with what b knows." (Hintikka 1975: 28) So, the knowledge of a person refers to a class of his or her knowledge worlds. These worlds are characterized relative to an agent, and because of this Hintikka (1973: 6) calls them "personal modalities." However, this does not imply any kind of subjectivity because the structure of the class of possible

worlds is determined conceptually. Moreover, the example shows that the modal notions are "local notions." We are not considering all the possible worlds but some well specified classes of possible worlds.

The spectrum of modal notions allows us to generate the conceptual framework context sensitively, and the "thought experiments" related to possible worlds can be specified carefully such that they are conceptually "clear and distinct." So, for example, Wilkes's examples do not characterize the actual situation; Wilkes (1988) connects the water and H_2O relationship to logical possibility, but this need not be the case. See, for example, Chang 2012.

The problem of identification is not merely that we must specify the modal context carefully, but also that we have to unify several different modalities together. For example, a person may know about him or herself. This knowledge should unify his or her beliefs about him and herself, but also hopes, for example, that others see him or her as something. In these kinds of situations, these different modalities may contradict each other and hence, for conceptual reasons, cannot be unified. However, our method of analysis cannot "solve" the personal problem, but it can clarify the situation.

6. Closing Words

The notion of identification is extremely complex and it has been used in several different fields of science. However, it is reasonable to separate the notion of identification from the notion of identity. Identification can be understood in several different ways. The detective is trying to identify the murderer so that he or she can distinguish the appropriate individual from the population independently, whether that person may be Dr. Jekyll or Mr. Hyde. The judge in the court room has to decide whether Dr. Jekyll is judicially responsible for the murder performed by Mr. Hyde. The medical doctor tries to identify whether there is one or two personalities within the same body. On the one hand, philosophers give us methodological tools for all these different professions to perform their identification tasks excellently – which refers to both ethical and professional excellence. Moreover, philosophers are still faced with the metaphysical problem of identity, which is a huge philosophical problem. The understanding generated by answering the metaphysical problem must be used in the methodological analysis of identification.

REFERENCES

Ayer, A. J. (1976). *The Central Questions of Philosophy*. Penguin Books.

Boer, S. E. & Lycan, W. G. (1986). *Knowing Who*. Cambridge: Bradford Book.

Carnap, R. (1967). *The Logical Structure of the World/Pseudoproblems in Philosophy*. University of California Press.

Chang, H. (2012). *Is Water H_2O? Evidence, Realism and Pluralism*. Dordrecht: Springer.

Descartes, R. (1641/1980). *Discourse on Method and Meditations of First Philosophy*. Hackett Publishing Company.

Gleason, P. (1983). "Identifying Identity: A Semantic History." *The Journal of American History* Vol. 69 (4): 910-931.

Hall, S. (1996). Introduction: "Who Needs 'Identity'?" In S. Hall and P. du Gay (eds.) *Questions of Cultural Identity*. Sage Publications.

Hendricks, V. F. (2001). *The convergence of scientific knowledge: a view from the limit*. Dordrecht: Kluwer Academic Publishers.

Hintikka, J. (1962/2005). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press.

Hintikka, J. (1963). "Cogito Ergo Sum as an Inference and a Performance." *Philosophical Review* 72: 487-97.

Hintikka, J. (1969). *Models for Modalities: Selected Essays*. Dordrecht: D. Reidel Publishing Company.

Hintikka, J. (1973). *Logic, Language Games, and Information: Kantian Themes in the Philosophy of Logic*. Oxford: Clarendon Press.

Hintikka, J. (1975). *The Intentions of Intentionality and Other New Models for Modalities*. Dordrecht: D. Reidel Publishing Company.

Hintikka, J. and Hintikka, M. (1989). *Logic of Epistemology and the Epistemology of Logic*. Dordrecht: Kluwer Academic.

Kripke, S. A. (1972/1980). *Naming and Necessity*. Harvard University Press.

Lewis, D. (1973). *Counterfactuals*. Oxford: Basil Blackwell.

Mach, E. (1976). "On Thought Experiments." In *Knowledge and Error*. Dordrecht: D. Reidel Publishing Company.

Nietzsche, F. (1887/2006). *On the Genealogy of Morality*. Cambridge University Press.

Nietzsche, F. (1988). *Ecce Homo: How One Becomes What One Is*. London: Penguin Books.

Olson, E. T. (2006). "Imperfect Identity." *Proceedings of the Aristotelian Society* 104: 81-98.

Olson, E. T. (2015). "Personal Identity." *Stanford Encyclopedia of Philosophy*. Read 13.6.2016.

Perry, J., Bratman, M. and Fischer, M. J. (eds.) (2010). *Introduction to Philosophy: Classical and Contemporary Readings*. Oxford University Press.

Pihlström, S. (2015). "Controlling Death: Philosophical Thanatology Meets Pragmatism." *Mortality* Vol. 20 (1): 48-66.

Quine, W. V. O. (1952). *Methods of Logic*. London: Routledge and Kegan Paul.

Quine, W. V. O. (1953/1980). *From a Logical Point of View*. Harvard University Press.

Quine, W. V. O. (1960). *Word and Object*. Cambridge: The MIT Press.

Quine, W. V. O. (1969). *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Quine, W. V. O. (1970). *Philosophy of Logic*. Harvard University Press.

Russell, B. (1912/1985). *The Problems of Philosophy*. Oxford University Press.

Russell, B. (1918/1988). *The Philosophy of Logical Atomism*. Open Court.

Ryle, G. (1949/2000). *The Concept of Mind*. The University of Chicago Press.

Sartre, J.-P. (1956). *Being and Nothingness: A Phenomenal Essay on Ontology*. New York: Washington Square Press.

Sartre, J.-P. (1968). *Search for a Method*. New York: Vintage Books.

Wilkes, K. V. (1988). *Real People: Personal Identity without Thought Experiments*. Oxford: Clarendon Press.

17. Meta-Representational *Me*

TAKASHI YAGISAWA

1. Introduction

The topic of this paper is the notion of the first person (singular), namely the notion *me*. Let us begin by distinguishing it from a different notion which is often confused with it, namely the notion *self*.

The notion *me* applies to me and me alone absolutely, whereas the notion *self* applies to me relative to me, applies to you relative to you, applies to Jill relative to Jill, applies to Jack relative to Jack, and so on. Everyone is the self relative to her/him; for every x, x is the self to x.[1] But only I am me, period. Of course, you may assert correctly, "Only I am me." But the content of your assertion when you say this does not deal in the notion *me*; for your word "me" does not express the notion *me*. Only my word "me" does. It is not even that your word "me" expresses the notion *me* to you. To you your word "me" expresses a certain notion, which you call "the notion *me*." But what you call "the notion *me*" is not the notion *me*, any more than the person you call "me" is me.

It might be suggested that the notion *me* is reducible to the notion *self* in the following way:

> Start with "For any x and any y, x is self to y iff x is y" and then let y be me. The result is "For any x, x is self to me iff x is me," which may be understood as the definition of "x is *me*" as "x is self to me."

This suggestion for reduction does not succeed, for the alleged *definiens* "x is self to me" contains word "me," which expresses the notion *me*. Replacing "me" in the *definiens* with "Takashi Yagisawa" will not do, for "x is self to Takashi Yagisawa iff x is *me*" cannot be said even by me to be true by definition, as I might not know that Takashi Yagisawa is me.

---

[1] If for every x, the relation of selfhood relates x to x and to nothing else, then the relation of selfhood is indistinguishable from the relation of identity. I do not see any harm in this, but if one wishes to maintain that selfhood applies only to individuals capable of conscious awareness of some sort, one may restrict the range of the variable "x" to such individuals.

The notion *me* is expressed by the word "I," which belongs to a semantic category David Kaplan calls *pure indexical*. (Kaplan 1989) This category also includes the words, "now," "here," and "actual." They are called "indexicals" because their extensions are determined by contextual factors. They are called "pure" because the determination is achieved without reliance on any specific action (e.g., pointing) or intention (e.g., referential intention) of the agent (speaker, writer).

Let us review the best Kaplanian indexical theory of meaning for "I" briefly and note some shortcomings of that theory viewed from the perspective of someone who is interested in explicating the notion *me*. I shall then propose an explication which overcomes the shortcomings. The explication articulates the logical origin of the notion *me* as a certain way things are represented. Thus, our investigation will start with semantics and proceed to conceptual explication.

2. Indexical Theory

The standard Kaplanian indexical theory of "I" is a broadly Fregean theory of meaning, according to which meaning determines reference.[1] The "indexical" part of the theory introduces a theoretical machinery called a "context of utterance," which is an ordered n-tuple, where $n \geq 4$. A *minimal context of utterance* is an ordered quadruple (4-tuple), $<a, t, p, w>$, where a is an individual (agent), t is a time, p is a place, and w is a world, where a is at t located in p at w. $<a, t, p, w>$ is minimal in the sense that any ordered n-tuple that is a context of utterance must include these four members in its initial segment. The four members correspond to the four indexical words, "I," "now," "here," and "actual," respectively. These words are said to be pure indexicals because their reference is completely determined by a minimal context of utterance.

A sentence type is said to express a proposition relative to a context of utterance. Note that expression is a three-place relation among a sentence type, a proposition, and a context of utterance, rather than a two-place relation between a sentence token and a proposition. The proposition expressed by a sentence type relative to a context of utterance will be subject to truth-value evaluation with respect to a *circumstance of evaluation*. If the sentence contains the word "I" ("me," "my," or "mine"), the word refers to the agent of the context of utterance and the expressed proposition concerns that individual, irrespective of who the agent is in the circumstance of evaluation. That is, the agent of the context of utterance relative to which

---

[1] In my exposition I depart from Kaplan's original terminology slightly for simplicity's sake.

a particular proposition is expressed by the sentence in question is the referent of the word "I" as it occurs in the sentence, and in the evaluation of the proposition in question for its truth-value with respect to any circumstance of evaluation, the original agent of the context of utterance figures in the same crucial way, no matter what individuals are involved in whatever manner in the circumstance of evaluation.

The indexical theory of "I" has many virtues, one of which is its well-known capacity to explain the contingent but *a priori* nature of

(1) I am here now.

The separation of a context of utterance and a circumstance of evaluation allows a discrepancy between the spatiotemporal location of the individual who is the agent in a context of utterance and the same individual's spatiotemporal location in a circumstance of evaluation, which explains the contingency. As for the apriority, the theory explains it by making it the case that in whatever context of utterance relative to which (1) expresses a proposition, that proposition is true relative to any circumstance of evaluation with the same agent, time, place, and world as the context of utterance.[2]

Like (1), the following (2) also does not express a necessarily true proposition:

(2) I am making an utterance here now.

Even if I am in fact making some utterance here now, there is no necessity about my doing so; I could easily be silent here now. But is (2) not a truth of indexical logic just like (1)? That is, is (2) not true with respect to every ordered n-tuple as regarded simultaneously as a context of utterance and as a circumstance of truth-value evaluation? The answer is "No." For some n-tuple that is a context of utterance, no utterance is made by the first member (agent) at all — the label "agent" is just a label, nothing more— hence no utterance needs to be made by the first member at the second member (time) in the third member (place) at the fourth member (world). The negation of (2) is indexical-logically coherent. Indexical logic does not treat the predicate "utter" ("make an utterance") as a logical constant, for the notion of uttering—unlike the technical notion of context of utterance (ordered n-tuple)—is not an indexical-logical notion. This is where the *naïveté* of Hans Reichanbach's phrase "token-reflexive" in his pioneering

---

[2] Some might question the apriority of (1) on the basis of examples such as the announcement, "I am not here now," as part of an outgoing message on a telephone answering machine. But such examples only highlight a temporal gap between the time of production of a sentence token (the time of the recording of the message) and the time of utterance (the time of the playback of the message).

work looms large. (Reichenbach 1947) The Kaplanian indexical theory sheds this *naïveté* and deals with expression types instead of tokens.[3]

Another virtue of the indexical theory is that it clearly distinguishes "I" and "self" in the way noted in the opening section. "I" refers to the agent in the context of utterance (the first member of the ordered quadruple), and this reference does not vary from one circumstance of evaluation to another. By contrast, "self" is not even a referring expression; it usually occurs as fused to a pronoun (e.g., "herself"). For example, "the speaker herself" denotes whoever is uniquely speaking in a given circumstance of evaluation, independently of who is speaking in a given context of utterance. When the accompanying singular term is a proper name or a pronoun, the context of utterance determines reference, but the reference need not be to the first person; "Jill herself" refers to Jill, and "she herself" refers to the female person most saliently featured in the context of utterance.

The indexical theory is quite attractive, and with its rigorous formulation as a formal semantic theory of the linguistic meaning of "I," it is difficult to find a serious defect. But as a philosopher, I am interested in more than just the linguistic meaning of the word "I" and its formal semantics. I am also interested in the notion *me*, which the word expresses. Providing a satisfactory formal semantic theory of the word "I" is one thing; elucidating the notion *me* satisfactorily is another. Impressive and useful as is, the indexical theory of "I" is not quite sufficient for giving a fully satisfactory explication of the notion *me*.

3. Shortcomings of the Indexical Theory

Let us consider what is lacking in the indexical theory when regarded as an explication of the notion *me*. Even though the Kaplanian indexical theory sheds the *naïveté* of Reichenbach's theory, it still has three important shortcomings.

First, if someone who is unfamiliar with English is told that "I" refers to the agent in the context of utterance, she will not be given sufficient information for associating the notion *me* with the word "I."[4] Will it help if she is told more fully that "I" refers to x relative to any context of utterance in which x is the agent? No, it will not. Remember that according to the formal semantic theory of indexicals, a context of utterance is simply an ordered n-tuple, where the first member is located at the other three mem-

---

[3] By contrast, the sentence, "The one who is making this utterance here now is making an utterance here now," is a truth of indexical logic, unless the reference of "here" or "now" is allowed to vary within a single sentence. (see Yagisawa 1993)

[4] Saul Kripke makes a similar complaint in Kripke (2011b).

bers. The first of the members included in the *n*-tuple is an individual, and the label "agent" is attached to that individual as such. Do not forget that for the formal semantic purposes, the label is a mere tag without connotation beyond "the first member of the *n*-tuple."

Second, the indexical theory makes the meaning of "I" invariable from speaker to speaker. Whether I use the word "I" or you use the word "I" or Jill uses the word "I," the meaning of the word "I" as so used remains constant. This is what we should expect of the meaning of "I" and linguistic meanings in general. But this clearly fails to capture the uniqueness of the notion *me* noted in the opening section. I can use the word "I" to refer to me, you can use the same word "I" with the same meaning to refer to you, and Jill can use it in the same way to refer to Jill. But the notion *me* applies to me and only to me, not to you or to Jill or to anyone else. The notion you express by your word "me" is not the notion *me* (the notion I express by my word "me"), even though the meaning of the word "I" ("me") you use is the same as the meaning of the word "I" ("me") I use.

Third and most importantly for our purposes in this paper, rigidity is left unexplained. The key idea of the indexical theory, as we have seen, is that the reference of the word "I" is relative to a context of utterance, and for any context of utterance *c*, "I" refers to the agent in *c*. This reference is rigid in the sense that it remains constant (to the agent in *c*) with respect to any circumstance of evaluation. This is as it should be and perfectly acceptable as the basic idea anchoring a formal semantic theory of "I." But it falls short of laying out the notion *me* in an informative way. The indexical theory, as it were, simply declares "I" to be a rigid designator. Distinguishing propositional expression relative to a context of utterance and truth-value evaluation with respect to a circumstance of evaluation enables the theory to treat "I" as rigid, and the theory goes ahead to treat "I" as rigid. But the theory offers no explanation why "I" should be treated as rigid in the first place.

The original categories of linguistic expressions on which Saul A. Kripke, who introduced the notion of rigidity, focused in introducing the notion of rigidity were proper names and natural-kind terms. Proper names and natural-kind terms lack descriptive meanings—or so Kripke (1972) argued. Kripke also sketched a causal picture of reference for these expressions, which is widely accepted by those who agree with Kripke on the non-descriptiveness of these expressions. But this picture of rigidity does not fit "I." It is not plausible at all to think that causation plays the same, or even similar, role in determining the reference of "I" as it is claimed to play in determining the reference of "Aristotle" or "tiger."

It is also unsatisfactory to say that proper names, natural-kind terms, and "I" are all rigid because they are all directly referential. "Is directly referential" is ambiguous between "refers without conceptual mediation" and "contributes the referent to the expressed proposition." The first understanding of direct reference plays a heavy role in Kripke's discussion, whereas the second is emphasized by Kaplan. The indexical theory is tailor-made to respect the direct referentiality of "I" in the second sense, but as we have seen, we want philosophical justification for respecting it. As for the first sense, it is not at all clear that "I" is directly referential in that sense. It may well be that the reference of "I" is determined by some non-trivial notion. My working hypothesis is that it is indeed so determined and that the notion me is that notion.

4. Representation

As already noted, the notion *me* is not, and does not entail, any linguistic notion, including the notion of linguistic utterance. Kaplan's retreat from Reichenbach's idea of token-reflexivity by regarding the bearers of semantic values to be linguistic expression types rather than tokens is a move in the right direction. What is essential to the notion *me* is not any notion of linguistic act but the notion of cognitive act, i.e., act of entertaining a content. The Cartesian conception of the first person as *res cogitans* (thinking thing, i.e., thought-content-entertaining thing) comes close. The Cartesian *cogito* argument is a good example of a traditional argument concerning the first person, which is internalistically driven. It starts by methodologically dispensing with the external world, and the ensuing solipsistic reasoning is supposed to suffice for the postulation of the Cartesian ego as an existent being. I propose instead that the first person be postulated in a way that is externalistically driven, in particular, that the notion *me* be regarded as having its conceptual origin in representation.

Representation involves three elements in addition to what does the representing: *content*, *object*, and *recipient*. Representation puts forth some content. Suppose that Jill has a visual experience of perceiving a spider descending from the ceiling by the door, that Jack has a qualitatively indistinguishable experience, and that Jill is actually perceiving a real spider, while Jack is merely hallucinating. What is common to their experience is the representational content. Jill's perceptual experience has a certain representational content, and Jack's perceptual experience has the same representational content.[5]

---

[5] Should someone think that the word "content" is inappropriate in view of the recent emergence of so-called *disjunctivism*, we might wish to choose a more neutral word.

Representation is about, or of, something, and that something is the object of representation. In the case of Jill's experience, there are a number of different objects of representation but the most prominent among them is the spider (the other objects of representation include the ceiling, the door, etc.). In the case of Jack's experience, even though there may be some objects of representation (e.g., the ceiling, the door, etc.), a spider is not one of them. There is a spider Jill is perceiving, but there is no spider Jack is perceiving.

Representation is to someone. Jill's perceptual experience has a certain content, which is put forth to Jill, and Jack's perceptual experience has a certain content, which is put forth to Jack. Perceptual experience without a perceiver is unintelligible. The perceiver is the recipient of the content represented by the perceptual experience. Generally, representation without a recipient is unintelligible.[6]

Should Jill fail to grasp the contrast between appearance and reality, she might be unable to distinguish one of the objects of representation (the spider) and a certain part of the content of representation (the "spider"-part of the "a-spider-is-descending-from-the-ceiling-by-the-door" content). Furthermore, if her self-consciousness were underdeveloped, she might not be able to realize that representation was occurring to some recipient.

5. *Me*-Way of Representation

Representation with the same content, with the same objects, and to the same recipient may occur in different ways. It may occur by the recipient's direct perceptual encounter with the objects, by her hearing about the objects, by her reading about the objects, or even by her dreaming about the objects. It may occur in such a way that the content is put forth clearly or in such a way that the content is put forth obscurely. It may occur in a way that is harmful to some of the objects or in a way that is beneficial to them. It may occur in a way that is threatening to the recipient or in a way that is inviting to the recipient.

Each of these and various other ways of representation divides into many overlapping sub-ways. For example, the content may be put forth clearly in different languages (English, Hungarian, etc.), in different font styles (block letters, cursive letters, etc.), at different decibel levels (loudly, quietly, etc.), in different emotional modes (angrily, calmly, etc.), through different kinds of behavior or tool (speech, writing, gesture, flag sema-

---

[6] A map of a city might be said to represent the city even when nobody is looking at it. But the map is intended to represent the city to whoever looks at it, so in that sense its representation requires a recipient.

phore, etc.), and so on. Cartographic representation of geographical information may be done by different methods of map projection (Albers Projection, Mercator Projection, etc.), in different colors (in blue ink, in red ink, etc.), and so on.

Among these and numerous other crisscrossing ways of representation is the *me*-way. What is the *me*-way? How is a given representational content put forth when the representation occurs in the *me*-way? Unlike the various example ways of representation mentioned above, which can be analyzed in more basic terms, the *me*-way is primitive, so no informative analysis can be offered. The notion "*me*-way" is not a compound notion made up of two more basic notions, "*me*" and "*way*"; I use an apparently syntactically compound noun phrase to designate the way for obvious expository reasons, but the notion expressed by the noun phrase is not conceptually compound. Conceptual analysis is not possible here. But I can be informative in other ways.

It might be helpful to compare my proposal with a well-known adverbial theory of perception by Roderick Chisholm (1957). According to Chisholm, when I see a red shirt, I sense the shirt in a particular way, namely *redly*. While Chisholm would say that I am sensing the shirt *redly*, I would say that the content of my perception is put forth in the *me*-way, or *me-ly*. For him, the shirt is sensed *redly*; for me, the entire content of visual perception is put forth *me-ly*. When the whole of my visual space is filled with the color red uniformly, Chisholm would say that I am appeared to *redly*, and I would say that a uniformly red content of visual perception is put forth *me-ly*. Chisholm would not speak of the visual content but would presuppose the notion *me* and predicate the property of being appeared to *redly* of me, whereas I do not presuppose the notion *me* but speak of the visual content and predicate the property of being put forth *me-ly* of it.[7]

If I directly saw a spider descending, then representation of the content that a spider is descending would be occurring in the *me*-way; the content that a spider is descending would be put forth *me-ly*. Suppose that as she directly sees a spider descending, Jill tells me that a spider is descending, while I keep my eyes closed. In such a case, the representation of the content that a spider is descending occurs with Jill as the recipient but not in the *me*-way. Representation can occur in the *me*-way only to me, that is,

---

[7] The point of comparing my proposal to Chisholm's theory is not ontological. Chisholm attempts to dispense with purely perceptual objects which are presumed to be the immediate objects of perception, such as sense data. He prefers to speak of ordinary physical objects as directly perceived, and I share his anti-phenomenalist ontological stance. But it is Chisholm's adverbial theory, in which he expresses ways of perception by adverbs, that I am drawing attention to.

only when the recipient is me. Jill might appropriately say, "Representation is occurring in the *me*-way." But what she calls "the *me*-way" is not the *me*-way, any more than she is me. At the same time, representation with a different content occurs in the *me*-way; a different content is put forth *me-ly*, namely the content that Jill is saying that a spider is descending.

When I speak of a way of representation, one might be reminded of Gottlob Frege's mode of presentation. (Frege 1980) Is a way of representation the same as Frege's mode of presentation? No. Frege's mode of presentation is associated with a linguistic expression and is identified as the sense of the expression, which determines the reference of the expression and the referent is the object presented. This forms the core of Frege's philosophy of language. I do not propose ways of representation as reference-determining linguistic properties of expressions, as should be clear from the examples I have given. Also, unlike Frege on his mode of presentation, I am not committed to the compositionality of ways of representation, namely, the principle that for any representation composed of sub-representations, the way of that representation is functional on the ways of the sub-representations plus how the sub-representations are put together to form the whole representation.

It is important to be clear about the relation between the *me*-way of representation and the notion *me*. The notion *me* applies only to me if it applies to anything at all, whereas there is no need for the *me*-way of representation to be a way of representing me. When I see a spider descending from the ceiling, the objects of the perceptual representation are the spider, the ceiling, etc., and not me. I do not figure in the content of the representation, which is put forth *me*-ly.

It would be a mistake, however, to infer from this that the *me*-way of representing the spider descending from the ceiling does nothing to help pick me out. It would be a mistake even to think that since all objects of representation in the *me*-way here are objects in the external world (the spider, the door, etc.), the *me*-way contributes nothing toward postulating me as an entity. If this were not a mistake, my proposal might be said to be Humean in some sense. But it is in fact not Humean in any sense; identifying the *me*-way as the conceptual origin of the notion *me* is not meant to be a step in an argument for Humean skepticism about the first person. By noticing and attending to the *me*-way of representation as I see the spider descending from the ceiling, I do succeed in picking out myself as the recipient of the representation. The *me*-way does successfully lead me to the notion *me*, hence the postulation of myself as an entity. I call this the

"way-to-thing shift,"[8] a transition from recognizing a way of representation to postulating the corresponding entity, the recipient of representation. This is done through extraction of the notion me from the *me*-way of representation.

6. Way-to-Thing Shift

The way-to-thing shift is a kind of phenomenon that is not unheard of elsewhere. For example, if a dancer and her partner move in tandem in a certain way, they may be said to be waltzing, and their dance a waltz. They move in a waltzing way, and they end up dancing a waltz. You might say that this is not an example of a way-to-thing shift, for a waltz is not a concrete thing. A waltz may not be a concrete thing, but "waltz" is a noun so that we have an example of an adverb-to-noun shift at least. Moreover a waltz may be defended as a thing on the ground that not all things are concrete and waltz is a non-concrete thing. Another example is a constellation. Betelgeuse, Rigel, Bellatrix, etc., are positioned in a certain way in the night sky seen from Earth. Because of that way of positioning, they are the constellation of Orion. It might be objected that a constellation is not a thing but a mere appearance of things. There is much to be said about the metaphysical status of constellations, but again we should note that "constellation" is a noun. Another example of a way-to-thing shift is a curve ball in baseball. A ball is thrown in a certain way, and because of the way it is thrown, it is a curve ball. Surely, a curve ball is a thing.[9]

Suppose that I see myself in a photograph and in it I am looking at a spider descending. Suppose that I fail to realize that the person in the photograph is myself. I understand that my visual perception of the photograph represents the situation of a spider descending in front of someone with a certain appearance, but fail to conceptualize that someone as myself.[10] What is the difference between this case and the case in which I realize that I am looking at myself in the photograph? How does the realization arise?

As my visual experience of the spider represents it as being thus-and-so to me, I may or may not realize that representation is occurring to someone. If I do realize it, I come to the realization by recognizing the *me*-way of representation as one of the ways in which the current representation is

[8] Or, to put it metalinguistically, the "adverb-to-noun shift."
[9] A curve ball is not an event, for one could not hit an event out of a ballpark.
[10] We might also say that my visual perception of the photograph represents the situation of there being a photograph of the scene of a spider descending in front of someone with a certain appearance. We might be able to point out other instances of the representation relation holding in this example, but we shall ignore them for simplicity's sake.

occurring. It is grasping this connection between the *me*-way and the presence of the recipient of representation that gives me the notion *me*.

Here, an analogy with time may be helpful. In particular, take the notion *now*.[11] We have been working under the assumption that representation is a many-place relation the *relata* of which are what does the representing, a content, an object(s), and a recipient. We may add one more *relatum*, a time. Assume that my perception of the spider is occurring at a particular time (moment or period) *t*. We may then say that my perceptual experience represents at *t*, with the content that a spider is descending, and the spider as the object and me as the recipient. It is important to note that the time *t* need not figure in the represented content as any specific time. Let *t* be noon on (a particular) Tuesday. I may not be aware of this, and my perceptual experience may not represent *t* as noon on Tuesday. That is, the content of representation may not be the "a-spider-descending-at-noon-on-Tuesday" content.

If I were yet to develop an explicit system of temporal notions, my perceptual experience would not put forth to me any content in which a particular time figured, so it would not only fail to put forth to me at *t* the content that a spider is descending at noon on Tuesday, but also fail to put forth to me at *t* the content that a spider is descending at *t*. Even so, the representation occurs in a certain way, which may be called the *now*-way. No matter how ignorant I may be of the temporal matters, as long as I experience the spider's descent as taking place at some particular time, my experience puts forth at *t* the "a-spider-is-descending" content *now-ly*.[12]

Contrast this with the case of my looking back later to *t* and remembering my experience at *t*. I will have a recollecting experience in which the "a-spider-is-descending" content is put forth not *now-ly* but *then-ly*.

---

[11] See Kamp 1971.

[12] Compare Kriegel 2015. Also, in footnote 68 on page 317 of Kripke 2011 (a), we read: "Suppose someone wonders what time it is now … So, in some sense, he is wondering what time it is, and the answer is given by the clock. Or he may be wondering when it will be noon, and the answer may be 'now', or 'two minutes from now'. 'When did she die?' 'Just now'. Both forms of question are legitimate, and equally so. In the first case, the very same situation is regarded in two ways. In my own opinion, the relativity and indeterminacy of '*wh*-questions' like this is exaggerated in the philosophical literature, but it exists and the present instance is a strong case." Kripke is saying that in the first case one and the same "situation" (what time it is now) is "regarded" in two different ways, which are such that depending on which way the "regarding" is understood, the correct answer is either (i) what the clock says, or (ii) how much temporal separation there is between now and noon. What I find important in this passage is that Kripke explicitly contrasts the "situation" and the way of "regarding" the situation; one and the same "situation" may be "regarded" in different ways. This distinction between a "situation" and a way of "regarding" it corresponds to my distinction between a content and a way of representing it.

As I become conscious of temporality of matters, I become able to entertain in the *now*-way not only the thought that a spider is descending but also the thought that a spider is descending now. When I do entertain the latter thought, the notion *now* both characterizes part of the represented content and is inherent in the way of the representation. Likewise, when I realize that I am looking at a spider descending, the notion *me* both characterizes part of the content of my perceptual representation and is inherent in the way of the representation; my perception represents not just the content that a spider is descending but also the content that I am looking at a spider descending (or the content that a spider is descending in front of me), in the *me*-way.[13]

7. Rigidity Explained

I see my own reflection on a large glass door at a pool party. Without realizing that I am seeing myself, I remain calm on the patio and say,

(3) His pants are on fire.

A moment later I come to the inevitable realization and jump into the pool, shouting,

(4) My pants are on fire![14]

The truth conditions of what is said by me using (3) and (4) are not different; what is said is true if and only if Yagisawa's pants are on fire. The reference of "his" and "my" is to the same man and equally rigid. But something important is different, for my behavior is importantly different. Obviously what is different is the crucial involvement of the notion *me* in the case with (4), which is absent in the case with (3). The rigidity of "I" ("my") is

---

[13] We may expand the analogy between *me* and *now* to include *actual*. David Lewis proposes an analysis of actuality in Lewis 1970 and distinguishes two senses of the word "actual," one rigid and the other non-rigid. It is customary simply to say, when discussing Lewis' analysis, that on that analysis every world is actual relative to itself: e.g., Leuenberger 2015: 111. When interpreted appropriately, such an analysis is correct for "actual" in the non-rigid sense only. The word "now" parallels the rigid sense of "actual," and for the non-rigid sense we may use the word "present": every time is present relative to itself. It is false to say that every time is now relative to itself. 1916 is not now in 1916. It is false even to say that 1916 was now in 1916. No time is, was, or will be now except *now*., viz., *this very time*. Likewise, nobody is me except *me*, and it is false to say that, say, Jack is me for Jack. The point may be made even more strongly if we replace "me" with "myself": nobody is myself except myself, and it is false to say that Jack is myself for Jack. The non-rigid sense is carried by just "self": Jack is the self for Jack, and everyone *x* is the self for *x*.

[14] This example is due to David Kaplan in Kaplan 1989. I have added the pool party setup.

explained if this involvement is explained. (In order to explain the rigidity of "his," one needs to offer a separate explanation.)

My perceptual experience in the case with (4)[15] represents the "Yagisawa's-pants-are-on-fire" content to me in the *me*-way. But this is not the crucial involvement of the notion *me* that distinguishes this case from the case with (3), for my perceptual experience in the case with (3) also represents the same content to me in the *me*-way; remember that all of my direct perceptual experiences represent contents to me in the *me*-way. The crucial involvement is the link between the *me*-way of representation and the "Yagisawa"-part of the represented content. The *me*-way of representation remains invariant from content to content; it is a way a given content is represented, rather than part of a given content. The *me*-way transcends particular contents in that sense and therefore has the effect of maintaining constancy through vagaries of differing circumstances of evaluation. Rigidity is none other than this effect. Once the "Yagisawa"-part of the represented content is linked to the *me*-way of representation, the rigidity effect kicks in. When I utter (3) calmly on the patio, the "Yagisawa"-part of the content is not linked to the *me*-way of representation. The rigidity effect kicks in only when the *me*-way of representation gives rise to the first-person conception of the recipient of the representation as a result of the way-to-thing shift, and also the recipient is identified as the individual corresponding to the "Yagisawa"-part of the content.

It might be objected that this explanation of the rigidity of "I" cannot be right because given the analogy with "now," if it is right it should also explain the rigidity of "now" but it does not. The reason why it might be said to fail to explain the rigidity of "now" is that once the way-to-thing shift occurs, *now* immediately gets incorporated into the represented content: "a-spider-is-descending" becomes "a-spider-is-descending-now." The objection might be fortified by the observation that the content without the *now* element can be a temporally neutral content whose truth-value is evaluable relative to different times with possibly different results, whereas the content without the *me* element cannot be a recipient-neutral content whose truth-value is evaluable relative to different recipients with possibly different results.

This objection misunderstands the proposed explanation. The proposed explanation acknowledges the disanalogy in question between *me* and *now* but only claims to explain the rigidity of "I" on the basis of the understanding of the notion *me* as primarily a meta-representational notion. Since *now* is also primarily a meta-representational notion, the same explanation

---

[15] The crucial mode of perception is unlikely to be visual.

of the rigidity of "now" applies. The disanalogy between "I" and "now" is not relevant to the explanation of the rigidity of either term.

The rigidity of "I" has quite a different basis from the rigidity of a proper name or a natural-kind term. The rigidity of a proper name or a natural-kind term arises from the semantic fact that such an expression has precisely the semantic role of inserting its referent into the propositions expressed by sentences containing it. The rigidity of "I" arises from the conceptual origin of the notion *me* in the *me*-way of representation, which makes the notion *me* transcend the represented contents, or the propositions expressed by sentences containing the word "I," and therefore the represented contents cannot shift from one circumstance of evaluation to another, as long as I am the one who is evaluating the content. The evaluator remains constant, for he remains to be I, as different circumstances are considered for evaluation, and since the referent of "I" is the evaluator (as the joint result of the way-to-thing shift and the infusion of the recipient of representation into the represented content), the referent of "I" remains constant. In this sense, the rigidity of "I" is meta-semantic in origin.

It should be remembered that I am not proposing a rival to the indexical theory. Nor am I casting doubt on the adequacy of the indexical theory as a semantic theory of "I," as far as it goes as intended by its proponents. The linguistic meaning of "I," like the meanings of other indexical terms, is shared by many speakers, and the indexical theory captures that meaning well. My concern is not linguistic but conceptual. My proposal is aimed at capturing the nature of the notion *me*, rather than capturing the linguistic meaning of the word "I." In fact, my proposal explicitly denies the public shareability of the notion *me*, and therefore misses the publicness of the linguistic meaning of the word "I." This, however, does not mean that my proposal is incompatible with a satisfactory theory of the linguistic meaning of "I."

What you express by the word "me" stands in a parallel (the same?) relation to you as the relation in which the notion *me*—which I express by the word "me"—stands to me. This might be sufficient for making my proposal adequate as a basis for an acceptable theory of linguistic meaning. In order to serves as such a basis, my proposal needs to be augmented by a theory linking the notion *me* and the notion you associate with the world "me" on one hand and the publicly shareable linguistic meaning of the word "me" on the other.[16]

---

[16] I thank Naoya Fujikawa for pushing me to clarify this point.

I hope that it goes without saying how my proposal overcomes the other two shortcomings of the indexical theory noted in section II. I leave it to the reader spell out how it its done.[17]

REFERENCES

Chisholm, R. (1957). *Perceiving: A Philosophical Study*. Ithaca: Cornell University Press.

Frege, G. (1980). "On Sense and Reference." In Geach, P. & Black M. (eds.) *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: Blackwell, third edition.

Kamp, H. (1971). "Formal Properties of 'Now'." *Theoria* 37: 227-274.

Kaplan, D. (1989). "Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals." In Almog, J., Perry, J. & Wettstein H. (eds) *Themes from Kaplan*. Oxford University Press: 481–563.

Kriegel, U. (2015). "Experiencing the Present." *Analysis* 75-3: 407-413.

Kripke, S. A. (1972). "Naming and Necessity." In Davidson, D. & Harman G. (eds.) *Semantics of Natural Language*. Dordrecht: D. Reidel: 253-355, (763-769); later published as a book with the same title from Harvard University Press (1980).

Kripke, S. A. (2011a). "Frege's Theory of Sense and Reference: Some Exegetical Notes." *Philosophical Troubles: Collected Papers.* Vol. I. Oxford University Press: 254-291.

Kripke, S. A. (2011b). "The First Person." *Philosophical Troubles: Collected Papers*. Vol. I. Oxford University Press: 292-321.

Leuenberger, S. (2015). "The Contingency of Contingency." *The Journal of Philosophy* CXII-2: 84-112

Lewis, D. (1970). "Anselm and Actuality." *Noûs* 4: 175-188, included with *Postscripts* in Lewis' *Philosophical Papers*, Vol I. Oxford University Press, 1983: 10-25.

Reichenbach, H. (1947). *Elements of Symbolic Logic*. New York: Free Press.

Yagisawa, T. (1993). "Logic Purified." *Noûs* 27: 470-86.

INDEX OF NAMES

Can we upload ourselves to computers? Are our iPhones and Notebooks parts of us? Why is *Know thyself!* still a good advice? What did the logical positivists think about the *Cogito*? Why did the Buddha argued that the self does not exist? Did the Ancients have a notion of subjectivity? What does it mean to put oneself in someone else's shoes? Should we feel sorry for those who have never been born? In what sense does Emma Bovary exist? What counts as a good answer to the question "Who am I?"

In this collection of articles an international team of seventeen authors raise and answer these and many other interesting questions about the self and related matters.

———————————————

Boran Berčić is Professor of Philosophy at the University of Rijeka, Croatia.

———————————————

CONTRIBUTORS

Eric T. Olson, Miljana Milojević,
Zdenka Brzović, Nenad Miščević,
Luca Malatesti, Boran Berčić,
Ljudevit Hanžek, Goran Kardaš,
Ana Gavran Miloš, Matej Sušnik,
Filip Čeč, Marko Jurjako,
Marin Biondić, Iris Vidmar,
Marta Ujvari, Arto Mutanen,
Takashi Yagisawa