BELGRADE PHILOSOPHICAL ANNUAL/
FILOZOFSKI GODIŠNJAK 36(1)/2023
Institute of Philosophy, Faculty of Philosophy,
University of Belgrade
Belgrade, Čika Ljubina 18–20

Belgrade

Year XXXVI

YU ISSN 0353-3891

UDK-1

Belgrade Philosophical Annual is published twice a year and is available online at http: //www.f.bg.ac.rs/bpa/index.html

Printed by
Službeni glasnik, Belgrade

This issue is financially supported by
Ministry of Education, Science and Technological Development of the Republic of Serbia

The statement on publication ethics can be found at the journal website http://www.f.bg.ac.rs/bpa

BELGRADE PHILOSOPHICAL ANNUAL 36(1)/2023

PERSONAL IDENTITY

PERSONAL IDENTITY

Eric T. Olson
e.olson@sheffield.ac.uk

Original Scientific Paper
UDC: 165.242.1-055.76
159.923:165.12

PARTIAL TWINNING AND THE BOUNDARIES OF A PERSON

Abstract: *In special cases of partial twinning, two heads, each supporting a more-or-less normal human mental life, emerge from a single torso. It is often argued that there must be two people in such a case, even if there is only one biological organism. That would pose a problem for 'animalism', the view that people are organisms. The paper argues that it is very hard to say what sort of non-organisms the people in such cases would be. Reflection on partial twinning is no more comfortable for those who think we're not organisms than for those who think we are. We may have to accept that a single person could have two separate mental lives.*

Keywords: *Animalism, constitution view, embodied mind account, personal identity*

1. The metaphysical puzzle of partial twinning

In special cases of partial twinning, two heads emerge from a single torso: *dicephalic parapagus* in medical jargon. Each head may contain a normal brain. The best-known example is the Hensel twins, who have not only survived to adulthood but passed their driving test and graduated from university. Between them they have two hearts, two stomachs, four lungs, and two spinal cords merging at the tailbone, but only one liver and intestine. Each brain controls the limbs on its own side and not those on the other.

How many people are there in such a case—two, sharing their lower parts, or just one with extra organs, like someone with a third kidney? Are there two thinking, conscious beings with one head each, or one conscious being with two heads? There might of course be both: two partially overlapping twins and also a two-headed person having both twins as parts. I'll set this unattractive suggestion aside. But there can't be both two people and also just one.

It's very natural to say that there are two. Tabloid papers may speak of two-headed babies, but everyone else describes Abigail and Brittany Hensel as twins, as I myself did a moment ago. This is not just because we see two faces and hear two voices, though that's of course a completely reliable indication of there being two people in all cases that most of us will ever encounter.

 DOI: 10.5937/BPA2336007O

The deeper and more important fact is that each face manifests a separate mental life: a set of sensations, attitudes, emotions, and so on that interact in a characteristic way with each other and not with anything outside the set. It appears as if the owner of one face can be awake while the owner of the other is asleep, and that they can have different beliefs, plans, preferences, and experiences. Call this the *two-person view*.

The alternative—the *one-person view*—is that there is just one person, with two brains and a radically disunified mental life—or, if you like, with two independent mental lives.[1] She could have blatantly inconsistent beliefs, plans, and preferences, but because they're realized in different brains, this would not cause any cognitive dissonance as it would in you or me. She could even be both awake and asleep at once. That's not to say that she could be both awake and not awake. She would be conscious in one brain, so to speak, and unconscious in the other.[2] This is a lot more counterintuitive than the two-person view. We know, more or less, how to interact with people who share parts. We don't know how to interact with someone who has two separate mental lives. It's easier, on the face of it at least, to suppose that there are two psychologically ordinary people who are physically joined than one person who acts like two.

No one would say that there is just one person in *every* case of partial twinning. There are doubtless two in cases where there are two torsos, for example, with four arms, joined to a single pair of legs. The serious one-person view is that one person *could* have two mental lives owing to partial twinning, and the serious two-person view is that there must be two people wherever there are two normally functioning brains.

If the one-person view is so implausible, why would anyone hold it? Well, because people are animals: biological organisms of the animal kingdom. And even if most cases of dicephalic parapagus involve two overlapping animals, there may be cases where there is only one, with two brains. Given that all people are animals, there cannot be one animal but two people. There must be either one person or none. And no one would deny that there is anyone there at all. That leads to the one-person view.

This shows that 'one-personers' and 'two-personers' disagree about deeper matters than just the number of people in these unusual cases. One-personers think we're animals. (Someone could hold the one-person view for another reason, without taking us to be animals, but I'm not aware of anyone who does or of any such reason.) Two-personers don't think we're

1 Some say she would have 'a divided mind'. I dislike this phrase because it suggests that there are things called 'minds'—a term that no one ever defines. When people speak of 'minds' as countable objects, I often don't know what they're talking about.

2 I discuss arguments for the two-person view and defend the logical consistency of the one-person view in Olson 2014.

animals. If they have any view on the matter, they think we're *not* animals. When philosophers discuss the metaphysics of partial twinning, it's almost always in order to argue for this claim.[3] Their reasoning is this: if there could be two people but just one animal in such a case, those people could not both be animals. (If there is just one F and there are two Gs, it follows that at least one G is not an F.) Presumably neither person would be an animal, as both would relate to the animal in the same way. And if the people in these cases would not be animals, no human people are. (I'll revisit this last step in the next section.)

The two sides also disagree on the more general question of what determines how many people or thinking beings there are at any one time. The reason for supposing that there must be two in these cases is that there are two separate mental lives: two systems of beliefs, preferences, plans, sensations, memories, and other mental items, each of which is in some sense unified. The items within each system relate to each other as your mental states and activities relate to each other, and they relate to those in another system as your mental states and activities relate to mine. More generally, two-personers take the number of people existing at a given time to be determined by facts about psychological unity and disunity.

One-personers deny this. They say that the number of people existing at a given time has nothing to do with psychological unity and disunity. They may accept that to be a person, as opposed to a nonperson, is a matter of psychology: to have certain special mental properties such as rationality and self-consciousness. (What it is to be a person, as opposed to a nonperson, is different from the question of what determines how many people there are.) But they take the number of human people to be necessarily equal to the number of human animals having those special mental properties, because human people *are* animals. Given the apparent fact—which two-personers do not dispute—that the number of animals is not determined facts about psychological unity, it follows that the number of human people existing at a given time is not determined by such facts either.

I concede that the attraction of the two-person view is a reason to deny that we are animals. But whether it's a *good* reason depends on whether that view is compatible with any plausible alternative account of what we are. If we're not animals, we must be nonanimals of some sort. And it's very hard to give a detailed and plausible account of what we are that is compatible with the two-person view. Reflection on partial twinning is no more comfortable for those who think we're not animals than for those who think we are. The one-person view may be the best of a bad lot.

3 E.g. McMahan 2002: 35–39, Gunnarsson 2010: 126–138, Campbell and McMahan 2016: 229–231. Campbell and McMahan take themselves to be arguing against the view that we are *essentially* animals, but this is a red herring: their argument would imply that we're not animals at all, essentially or otherwise.

2. Retrenched animalism

I'm not convinced that there is just one organism in any real case of dicephalic parapagus. I think the Hensel twins are probably two overlapping animals. But for all I know there *could* be a single organism with two brains, each supporting an independent mental life sophisticated enough to be that of a person. If people are animals, there can be only one person in such a case, with two independent mental lives.

I said the one-person view was motivated by the thought that we are animals: 'animalism'. Why suppose that we're animals? Well, it seems possible for a biological organism to think and to be conscious. And in that case the animal you see in the mirror is thinking and conscious right now. But it would be silly to suppose that it was a thinking, conscious being *other* than you. Wouldn't that make it a second person? And how could you ever know which of these people *you* were? To avoid these awkward questions, opponents of animalism must deny that human animals can think. More strongly, being a biological organism must be metaphysically incompatible with having any mental properties at all: if any organism could possibly have mental properties, the one you see in the mirror would now be thinking. This would be a metaphysical dualism of mind and life. Opponents of animalism face the unenviable task of explaining what it is about mental properties that makes them incompatible with being an animal. Why not accept the apparent fact that we're animals instead?

But even if *most* of us are animals, we may wonder whether the people in dicephalus cases must be. Might all people be animals *except* when there are more mental lives than animals? We might call this *retrenched animalism*.

Though it may have some superficial appeal, however, it's unlikely to satisfy anyone. All the objections to animalism apart from the one from conjoined twinning apply equally to retrenched animalism. They imply that *no* people are animals. It's often claimed, for example, that if your brain were transplanted into my head, and this gave the resulting person your memories, beliefs, preferences, and other mental properties rather than mine, he would be you and not me. More generally, people persist by virtue of some sort of psychological continuity. But no biological organism persists by virtue of psychological continuity. No animal would go with its transplanted brain: the operation would simply move an organ from one animal to another just as a liver transplant would. It would follow that even though no one will ever have such a operation, each person has a modal property that no animal has, namely persisting by virtue of psychological continuity. And anything having a property that no animal has is not an animal, ruling out even retrenched animalism.

Animalists won't be happy with it either. They'll want to know more about the thinking nonanimals in partial-twinning cases. What sort of

things would they be, and how would they relate to the animal? Why are there no such thinking nonanimals in ordinary cases—here in my chair right now, for instance? No animalist would accept that there is: otherwise it would be a second person in addition to me, raising the same 'too-many-thinkers' problem that features in the standard argument for animalism. And why is the human animal in a twinning case not itself a person—that is, not thinking and conscious? (If it were, there would be three people in such a case, not two.) I don't know any good answers to these questions. Retrenched animalism looks doomed to be a friendless view.

3. Partial twinning and the constitution view

The metaphysically interesting cases of partial twinning are those where a single animal has two normal brains: call them *one-animal-two-brain cases*. If there were two people in such a case, they would not be animals. What would they be, then? What sort of *non*animals? What properties would they have besides the mental properties that make them people? What parts would they have, if any, and where would their boundaries lie?

Now the same questions arise if ordinary people—those who aren't conjoined twins—are not animals. What sort of nonanimals are *they*? What nonmental properties do they have? The most common answer is that they're material things 'constituted by' animals: the 'constitution view'.[4] Each of us is made of the same matter as a certain animal, with the same physical properties and behaviour (in both actual and counterfactual circumstances), but different modal properties: we persist by virtue of psychological continuity; the animal does not. "Constitutionalists" generally say that differ from the animal mentally: otherwise the animal would itself be a person, one who *didn't* persist by virtue of psychological continuity. The usual constitution view is that strictly speaking animals have no mental properties at all, making them 'zombies' in the philosophical sense (Olson 2018).

But if there are two people in a one-animal-two-brain case, they can't be constituted by animals. If they were, they'd both be constituted by the *same* animal, seeing as there is only one animal there. They would each be made up of that animal's matter—all of it—and extend all the way out to the animal's skin. Each would have two heads. They would be physically indistinguishable, and that would presumably make them mentally indistinguishable as well: if the left brain is a part of both people, how could the mental activities going on in it belong to one person but not the other? Each would have two mental lives. But if it's hard to believe that there could be one person with two mental lives in such a case, it's even harder to believe that there could be *two* people with two mental lives each. It would be better to accept the one-person view.

4 E.g. Baker 2000, Johnston 2007, Shoemaker 2008.

This looks like an important objecton to the constitution view. No constitutionalist will be happy with the one-person view. That would mean accepting that there is never more than one person (at a given time) in any case where there is just one animal. The number of people at a given time would be determined by the number of biological organisms and have nothing to do with psychological unity and disunity. If you have to say that, you may as well accept that we're animals.[5] (That would also avoid the consequence that human animals are zombies.)

Could it be that all people are constituted by animals *except* those in one-animal-two-brain cases—'retrenched constitutionalism'? This would raise many of the same awkward questions facing retrenched animalism. If there is something constituted by an ordinary human animal, will there not be something constituted by a human animal having two brains? Will that thing not be a person with two mental lives? And if the two people in such a case are not constituted by animals, what *are* they constituted by? What are their parts and their boundaries? The constitution view tells us nothing about this.

This question about our boundaries arises not just on retrenched constitutionalism, but on any two-person view, or at least any who takes us to be material things. If some human people are smaller than a human animal, how big *is* a person? I'll devote the rest of the paper to this point.

4. The Lockean view of our boundaries

If there are two people in a one-animal case, they must be physically different (assuming they're physical things at all). They may share some of their parts, but they can't share all of them. And this physical difference must account for their mental difference: why each has just one mental life different from that of the other. It looks as if they must each have only one head and one brain.

What other parts would they have? Maybe the parts of the animal controlled exclusively by the right brain are parts of the twin whose brain it is—call her 'the right twin'—and of her alone, while those controlled exclusively by the left brain are parts of the left twin and of her alone. Or if 'control' means voluntary control, it's the people whose brains they are who have it. The left twin has exclusive voluntary control over the animal's left arm, for example, in that she can move it at will by means of a mental act realized in her brain (the animal's left brain), whereas the right twin alone can move the animal's right arm. And the reason why the animal's right head is not a part of the left twin may be that she has no voluntary control over

5 Some constitutionalists say that in a loose sense we *are* animals, even though each person is numerically distinct from the animal constituting her. But this description only confuses matters (Olson 2016) and does nothing to address any of the difficulties arising from partial twinning.

it. Any parts of the animal that both twins have voluntary control over are shared. This suggests a general principle about the parts of a person:

> Necessarily, a thing is a part of a person at a given time just if the person has voluntary control over it—she can move it at will—at that time.

(Or at least this would hold for any person who is a material thing. I won't speculate about what might make something a part of a god or an angel.)

This can't quite be right, as it implies that no atom is a part of anyone. No one has voluntary control over any single atom, yet we could hardly be made of matter unless the atoms making up that matter were parts of us. Perhaps atoms are parts of me by being parts of larger things that I *can* control: arms, say. In that case a thing is a part of a person just if she can either move it at will or move some larger thing that it's a part of. In other words, a person's parts are those things that she can control voluntarily and the parts of those things.

A related thought is that something is a part of a person because she can *feel* it. What makes an arm a part of me is my ability to feel sensations in it. And again, although I can't feel any individual atom, an atom might be a part of me by being a part of something I *can* feel. Locke once said something like this. The particles composing our bodies, he said,

> whilst vitally united to this same thinking conscious self, so that we feel when they are touch'd, and are affected by, and conscious of good or harm that happens to them, are a part of our *selves*: *i.e.* of our thinking conscious *self*. Thus the Limbs of his Body is to every one a part of *himself*: He sympathizes and is concerned for them. Cut off an hand, and thereby separate it from that consciousness, we had of its Heat, Cold, and other Affections; and it is then no longer a part of that which is *himself*, any more than the remotest part of Matter. (Locke 1975: 336f.)

Or we might combine these two thoughts: something is a part of me (a largish part) if I can both move *and* feel it, or if I can do one or the other. Call this the *Lockean view* of our boundaries. (Not to be confused with Lockean views of what it is to be a person, or of what it is for a person to persist through time. Locke held views on many questions in this area, and it's important to keep them separate.)

But the Lockean view faces grave difficulties. One is that it doesn't actually tell us where our boundaries lie. We can see this by noting that it does not imply that there must be two people in a one-animal-two-brain case: for all it says there may be just one, who can move and feel all the animal's limbs. It doesn't tell us why there must be two people rather than one, or why the animal is not itself a person. As an account of our boundaries it's radically incomplete at best.

More seriously, it appears to have the startling consequence that a limb that was numb and paralyzed could not be a part of anyone. Cutting the neural connections to your left leg would immediately make you smaller, giving you the size and shape of an amputee. Someone numb and paralyzed from the neck downwards would be literally a head. You might suppose that numb and paralyzed limbs are parts of someone because they're parts of a larger thing that she can move and feel, namely an organism. That would suggest that you can move and feel an organism if you can move and feel any part of it. It would follow that any person in a one-animal-two-brain case would have two heads, contrary to the two-person view. And whether you could move or feel something would be almost entirely irrelevant to whether it's a part of you, leaving the Lockean view with little content.

Finally, nerve damage could deprive you of the ability to move or feel anything at all. Any Lockean view will imply that this would result in your having no parts. But nothing can exist without having parts—even itself, its 'improper' part. It would follow that you could not exist in this condition: total paralysis would be instantly fatal. It would be metaphysically impossible for any conscious, intelligent being to be, even for a moment, entirely numb and paralyzed (apart perhaps from a god or an angel). No one will accept that.

5. The brain view

We've seen that if there can be two people in a one-animal-two-brain case, they cannot be animals. Nor can they be things constituted by animals, or things made up of those parts of the animal that a person can move at will or feel. What else might they be?

The obvious thought is that they're smaller parts of an animal: things like brains. And in that case you and I are brains too. The people in the twinning cases are brains only if brains are the entities in those cases that are conscious and intelligent—those satisfying the definition of 'person'. An animal (or a thing constituted by an animal) would be capable of thought only in the loose sense of having a brain-sized part that can think. What *really* thinks must be brains. Otherwise our being brains would imply that strictly speaking we *don't* think and are not conscious, but are merely parts of thinking beings, and no one will want to say that. And if it's brains that think in cases of partial twinning, it will be brains that think in ordinary cases too. So we must all be literally brains. Strictly speaking, each of us weighs about two pounds and is made up mostly of soft, yellowish-pink tissue.

I don't know of anyone who actually holds this view.[6] It would mean that we never really see or touch anyone, or even any part of anyone.

6 Some say that we're brains but don't mean it. Parfit says it, for example, yet in the same essay gives an account of our persistence that clearly does not apply to brains (Parfit 2012). Hudson (2001, 2007) thinks we're temporal parts of brains. His view avoids some but not all of the problems facing other versions of the brain view.

(Brain surgeons are an exception.) Nor could anyone walk or talk: except in an extended sense, those are not things that a brain can do. And despite appearances, conjoined twins would not actually share any parts. But I wouldn't worry too much about this. If I had a good reason to think that I was a brain, I could live with it. These are consequences of any view that denies that we're material things, yet almost no one objects to substance dualism (for example) on those grounds.

A more serious worry is that many parts of the brain are no more directly involved in producing our mental lives than our hearts and lungs are: the blood vessels supplying the nerve tissue, for example. If the blood vessels within the brain are parts of us, why not the blood vessels outside the brain? What's special about that boundary? The claim that we extend out to the surface of the cerebral cortex but no further looks arbitrary and unprincipled. At best a person might be a certain *part* of a brain, one excluding blood vessels and the like.

And the brain view is dialectically unstable. Suppose we're led to it by our allegiance to the two-person view. This reasoning is based on the thought that the number of people (or conscious beings generally) at a given time is determined by facts about psychological unity and disunity: about the number of mental lives in the sense characterized in §1. And that thought goes naturally with idea that the number of people existing during an extended period—the persistence of a person—consists in some sort of psychological continuity. To hold one of these views without the other would seem unprincipled, and I don't know of anyone who has ever done so. But this claim about our persistence is incompatible with our being brains, or even parts of them. Brains and their parts don't persist by virtue of psychological continuity. We could know whether your brain still exists without knowing anything about your mental properties. It could be removed from your head and pickled in a jar, just as your liver could, despite there being no psychological continuity in such a case. No one thinks that *you* could be removed from your head and pickled. If you were a brain, your persistence would have nothing to do with psychological continuity. But anyone who accepted that would take us to be animals rather than brains.

6. The embodied-mind account

Tim Campbell and Jeff McMahan propose a variant of the brain view intended to avoid these problems. They say:

> a person is identical to those functional areas of her brain that are necessary and sufficient for her capacity for consciousness.[7]

7 2016: 233. Let me emphasize that it's not my intention to single out Campbell and McMahan for criticism. If they make a number of problematic claims, that's because they actually try to answer the hard metaphysical questions that arise on the two-person view, rather than ignoring them as its adherents more commonly do.

This is meant to avoid the arbitrariness in the view that we're brains rather than larger things by saying that we are only those parts of the brain ('areas', as they say) that are necessary and sufficient for our capacity for consciousness. Presumably that excludes the brain's blood vessels. McMahan (2002: 66–94) calls this the 'embodied mind account'.

It requires a certain amount of explanation. First, it says that a person is certain *parts* of the brain. I don't think Campbell and McMahan meant to suggest that one thing could be numerically identical with many things—an idea of doubtful coherence at best (van Inwagen 1994, Sider 2007: 55–69). I assume they meant that a person is *composed* or *made up* of those functional brain parts that are necessary and sufficient for her capacity for consciousness. (Some things, the *x*s, compose something $y =_{df}$ each of the *x*s is a part of *y* and every part of *y* shares a part with one or more of the *x*s.) So each of us is a certain *part* (singular) of a brain.

Or rather a *functional* part.[8] The qualification is there to avoid the consequence that we could be pickled in a jar. The capacity for consciousness is essential to us: we couldn't exist without it (Campbell and McMahan 2016: 234). Because that capacity is not essential to a brain, we're not strictly brains, but 'functional brains': brains (or parts of them) having the capacity for consciousness. So when my brain stops functioning at death, the 'functional part' of it that I am does not simply stop functioning, but stops existing.

This, like one-many identity, may look like a logical confusion. It implies that the brain is one thing and the functional brain is another. That's like saying that a dog is one thing but a sleeping dog is something else. A dog can be either awake or asleep, but a *sleeping* dog can only be asleep: when it wakes up it doesn't just stop being asleep, but stops existing altogether, replaced by a waking dog that was never asleep. But surely a sleeping dog is not something other than a dog. To be a sleeping dog is simply to be both a dog and asleep. And of course a dog can survive the event of waking up. The sleeping dog exchanges the property of being a sleeping dog for the property of being a waking dog. That is, it remains a dog, loses the property of being asleep, and acquires that of being awake. That's what it is for a dog to wake up. And likewise, it would seem, something is a functioning brain just if it's both a brain and functioning. So when a brain stops functioning, it ought to exchange the property of being a functioning brain for that of being a nonfunctioning brain in the same way.

This is just what Campbell and McMahan deny. What then *is* a 'functioning brain'? How does it relate to nonfunctioning brain, or simply to a brain, without qualification? Someone might say that a functioning brain is something that the brain 'constitutes' until it stops functioning, at which point the functioning brain comes to an end but the brain itself carries

8 Similarly, Tye (2003: 143) proposes that we are 'brains insofar as they are in certain states', though he gives no account of what this means.

on—much as a lump of clay might constitute a statue but carry on after the statue is destroyed by squashing. But if the brain itself were thinking and conscious, that ought to make it a person—a second person in addition to the functioning brain—which, apart from being absurd, would contradict the embodied-mind view. Alternatively, if the brain itself were not conscious, it would be a zombie: a thing physically indistinguishable from an ordinary person but devoid of consciousness.

Campbell and McMahan avoid this awkward dilemma by saying that there is no such thing as 'the brain' without qualification. There is only the functioning brain, which is essentially functioning, and the nonfunctioning brain, which is essentially not functioning. The functioning brain stops existing when it stops functioning—that is, when it loses the capacity for consciousness—but it's not outlived by the brain *simpliciter*. Rather, a nonfunctioning brain comes into being to replace it. It's this nonfunctioning brain that can be pickled. The specimen in the anatomical museum labelled 'human brain' was never in anyone's head or able to support consciousness. Nor did my brain exist before it acquired that ability. There was of course a part of the foetus I grew from that anatomists call its brain, but it was not a brain in the sense that the organ in my head now is a brain. It was another nonfunctioning brain, which perished when it developed the capacity for consciousness. So a human organism starts out with a nonfunctioning brain, which is later replaced by a functioning brain that persists until it loses the capacity for consciousness, at which point it's replaced by another nonfunctioning brain. The person is the functioning brain (or rather a certain part of it).

Campbell and McMahan don't suppose that whenever anything loses *any* property, it thereby ceases to exist. They don't think that when a sleeping dog wakes up, it vanishes and is replaced by a waking dog that was never asleep. Their view is only about one special property, namely the capacity for consciousness.

7. Why we could not be embodied minds

What should we make of all this? It certainly goes against what anyone would otherwise think about the persistence of bodily organs. You'd be forgiven for thinking that an ordinary human being has just one brain, which comes into being in the course of foetal development, acquires in late gestation the capacity to produce consciousness, and at death loses that capacity and carries on for a time without it. On the embodied-mind account a human being has three brains, each existing during just one of these periods.

But I haven't got much faith in folk metaphysics. The real problems for the embodied-mind account lie elsewhere. It says, again, that a person is composed of those functional parts of her brain that are necessary and

sufficient for her capacity for consciousness. None of the the individual atoms composing a brain is 'functional'. The account is meant to tell us which larger things are parts of me. Some large parts of my brain are necessary for consciousness, in that I could not be conscious unless they were functioning, but not sufficient, in that their functioning alone would not make me conscious. The brainstem might be such a part. By contrast, the normal functioning of *all* my large brain parts is sufficient for consciousness but not necessary: I could be conscious even if some of them were disabled or destroyed. I am, then (according to the embodied-mind account) composed of those large brain parts whose functioning is individually necessary and jointly sufficient for me to be conscious. In other words, a large brain part is a part of me just if its functioning is necessary for me to be conscious and it's a member of a set of such parts whose functioning is jointly sufficient for it—and atoms are parts of me by being parts of one of these larger parts.

The trouble is that there is no such set. Consider my left cerebral hemisphere. Its functioning is not necessary for me to be conscious: I could be conscious without it. (Hemispherectomy—a real-life treatment for severe brain tumours—does not normally render the patient permanently unconscious.) So the embodied-mind account implies that my left hemisphere is not a part of me. My right hemisphere can't be a part of me either, for the same reason: it needn't function for me to be conscious. My only parts must be things that don't overlap either hemisphere: things like the brainstem, the amygdala, and the basal ganglia. But the functioning of such things is not sufficient for consciousness. To be conscious you need one hemisphere or the other, but not both. So any set of large brain parts whose functioning is jointly sufficient for consciousness will not be necessary for it, and any set of such parts whose functioning is necessary for consciousness will not be sufficient. The embodied-mind account entails that there are no people: you and I do not exist.

8. The supervenience-base account

Campbell and McMahan briefly mention a variant of the embodied-mind account saying that a person is 'the set of particles whose physical properties constitute the supervenience base for her phenomenal properties'—phenomenal properties being those characterizing the subjective nature of conscious experience.[9] What it's like for me to smell buttered toast in the

9 2016: 233. Hudson (2007: 219) makes a similar suggestion. The definition of 'phenomenal property' is my own. Campbell and McMahan define such properties as those a being has when it's in a state that there is something it's like to be in. That would make my current weight phenomenal, since I have it while being in a conscious state. Perhaps they meant a property that *necessarily* a thing has when *and only when* it's in such a state. The precise definition is unimportant for our purposes.

morning would be an example. I don't think Campbell and McMahan meant to say that a person is literally a set in the mathematical sense. Sets cannot have different members at different times, but a material person would be composed of different particles at different times. I assume they meant that a person is composed (at a given time) of those particles whose properties form the supervenience base for her phenomenal properties (at that time).

I suppose certain properties of some things, the *x*s, are the supervenience base for something *y* just if there can be no difference in *y* without a difference in those properties of the *x*s. So the thought is that a person is composed of those particles such that there can be no difference in her phenomenal properties without a difference in the physical properties of the particles. In other words, a person is made up of the particles that would have to be physically different for the subjective quality of her experience to be different. This will no doubt include not just the particles' intrinsic properties, but also the causal and spatiotemporal relations holding among them.[10] For example, I'm now having the auditory sensation of road traffic. (This is so even if I'm not actually hearing road traffic but only hallucinating: I'm speaking of the subjective quality of my experience.) For me to have the sensation of birdsong instead, certain particles would have to be differently arranged. More generally, there are certain particles that would have to be physically different for my current experience to be phenomenally different. Those particles are the ones that now compose me.

Call this the *supervenience-base account* of what we are. Though suggestive, I don't think it improves on the original embodied-mind account. There are probably no individual particles that would have to be physically different for my experience to be different. *Some* particles would have to be different—there couldn't be a difference in my experience without a physical difference in *any* particles—but there are no particular particles that would have to be different. Any individual particle could be the same, and the difference in my experience could be the result of differences in other particles. Imagine a picture composed of a million pixels. For it to look different, some of the pixels would have to be different: there couldn't be a difference in the visual appearance without a difference in any of the pixels. But there are no particular pixels that would have to be different: even if any individual pixel were exactly the same, the appearance might differ owing to differences in other pixels. It could be the same with the particles responsible for my experience. There appear to be no particles, the physical properties of which are the supervenience base for my phenomenal properties; and thus, on the supervenience-base account, there are no particles that compose me. Yet every material thing must be composed of particles. (Even an individual particle is, trivially, composed of itself.)

10 It might also include a specification of the laws of nature. Maybe the nature and arrangement of the particles produces a certain sort of conscious awareness only because of a contingent psychophysical law. I'll ignore this complication.

But set aside these tiresome matters. The idea behind the embodied-mind account is that a person is made up of just those particles that are directly involved in her experience. My hands and feet are not parts of me because, although they may be involved in my experience by contributing to my tactile sensations, they're not *directly* involved. Not all the parts of my brain are directly involved either. I'm made up of those that *are* directly involved. The story about supervenience bases is meant to spell out what 'direct involvement' is. It doesn't work, but there may be another that no one has yet thought of. Imagine that we have such a story.

9. The boundaries of conscious beings

As I understand it, the embodied-mind account is meant to be grounded in the nature of phenomenal properties. So it must hold not only for people: *any* conscious being must be made up, at a given time, of all and only the particles directly involved in its capacity for conscious experience at that time. It follows that dogs, fish, and pigeons cannot be conscious. At best they could be 'conscious' in the loose and derivative sense of having a part that's conscious strictly speaking—the same sense in which dogs are foot-shaped. It's metaphysically impossible for anything having feet or scales or wings as parts to be conscious. No biological organism (or for that matter anything 'constituted by' one) could have even the capacity for consciousness, because no such entity is made up entirely of particles directly involved in that capacity. And if an organism could never be conscious, it's hard to see how it could have any other mental property. This is another version of the metaphysical dualism of mind and life that I mentioned in §2.

If someone told me that conscious beings can never have feet as parts, I would expect her to say that this is because no material thing could have any mental property at all: conscious beings must be entirely immaterial.[11] That would be understandable: it really is hard to see how consciousness could arise out of physical activity. But I wouldn't expect a *materialist* to say it. The idea that a conscious being must be made up of all and only the physical particles directly involved in her conscious experience looks like an unhappy amalgam of Cartesianism and materialism.

Its consistency with the two-person view may perhaps give it some attraction. But it does not actually support that view. The embodied-mind account does not imply that there must be two conscious beings in a case of partial twinning, or even provide any reason to think so. For all it says, the particles directly involved in someone's consciousness might be spread across two independent brains. Campbell and McMahan's employment of it simply

11 Some materialists deny that there *are* any feet (van Inwagen 1990: 81–97), but the point can be made by replacing 'feet' with 'particles arranged footwise'.

assumes that in such cases the relevant brain parts in one head belong to one conscious being and those in the other head belong to another.

Of course, the embodied-mind account was not proposed as an argument for the two-person view, but as an answer to a pointed question arising from it, namely what sort of things those people are—more specifically, where their boundaries lie and why. But because it's consistent with the one-person view, the answer it gives to that question is far from complete. It says that every part of an organism that is directly involved in consciousness, and nothing else, is a part of a conscious being, but it says nothing about whether any two such parts belong to the same conscious being or different ones. It's entirely silent on the point of contention between the one-person and two-person views.

We could remedy this by combining it with the claim that nothing could be conscious unless its consciousness is in some way unified: no conscious being could have two mental lives in the sense spelled out in §1. But why should this be? Why could nothing have *dis*unified consciousness? That question must presumably have an answer, and the friends of the two-person view haven't given one.[12] Until they do, the embodied-mind account will not tell us what to say about partial twinning.

10. Totalitarian properties and the problem of mental specialists

Here is a final problem for the embodied-mind account—and for any verson of the brain view (Olson 2007: 87–98). The account says, again, that a conscious being must be composed of just those particles that are directly involved in her consciousness. But what about other mental properties? Is a being that remembers composed of just those particles that are directly involved in its remembering? The particles directly involved in my remembering are unlikely to be just those directly involved in my consciousness. There are almost certainly portions of my memory banks that would not have to be physically different for my conscious awareness to be phenomenally different. So if a remembering being had to be composed of just those particles directly involved in her memory, the embodied-mind account would entail that no remembering being was composed of the same particles as any conscious being. Nothing would both remember and be conscious. Seeing as *I*'m conscious—I could hardly be wrong about that—I'd be unable to remember anything. If I seem to remember things, that can only be because some other being—presumably another part of my brain overlapping with me—remembers, and its interaction with me somehow gives me the illusion that I'm remembering.

12 Shoemaker (1996), who thinks it follows from the functionalist theory of mind, may be an exception.

Or again, the particles directly involved in someone's thinking about philosophy are unlikely to be precisely those directly involved in her conscious awareness. If philosophers had to be composed of just those particles directly involved in their capacity for philosophical thinking, they could not be conscious. They'd be zombies. (It would be inappropriate for me to make a joke here.) And the same goes for other mental properties. What looks like a single being that is conscious, remembers, does philosophy, and so on would in fact be a large number of different beings—different parts of the brain—each having only a single mental ability. I take that to be absurd, and I'm sure the friends of the embodied-mind account would too. Call this the *problem of mental specialists*.

We can see that the original brain view faces it too. Why would anyone suppose that each of us is literally a brain—that we extend all the way out to the surface of the cerebral cortex and no further? The answer is presumably that a thinking being must be composed only of parts directly involved in her mental states and activities, or more generally in her mental properties. But which mental states, activities, or properties must someone's parts all be directly involved in? The particles directly involved in remembering will not be those directly involved in thinking about philosophy, recognizing shapes, and so on, even if there is considerable overlap. If every mental property must be had by a being whose parts are all and only those things directly involved in its having that property, then no being will ever have more than one mental property—because, again, different mental properties depend on different particles. This seems to me the most serious objection to any view on which we're brain-sized material things.

The problem would not arise if consciousness were metaphysically unique among mental properties. The embodied-mind account says that the capacity for consciousness can be had only by a being composed entirely of particles directly involved in it. That property forces direct involvement on every part of anything that has it. Like a totalitarian state, it demands not merely obedience, but active participation from every citizen around the clock. We might call such properties *totalitarian*. A property is totalitarian just if a thing can have it (at a given time) only if all its parts directly contribute (at that time) to its having that property. I don't know how to define the crucial phrase 'directly contribute to a thing's having a property', but any version of the brain view, as far as I can see, will require it. Let's see how far we can get without a definition.

We might wonder whether there *are* any totalitarian properties. *Being alive* in the biological sense might be one. Maybe all the parts of a living thing must be directly involved in its life processes: metabolism, immune activity, and so on (van Inwagen 1990: 81–97). It would follow that a prosthetic limb, a pacemaker, or a plastic heart valve—as well as any parts of an animal's hair and horns having no blood supply—could not be a part of an organism, but

only a part of its environment. *Being a material thing* might be another, given that all the parts of a material thing must themselves be material things.

The embodied-mind account says, again, that the capacity for consciousness is totalitarian: that's why a conscious being cannot extend beyond its brain. And any version of the brain view must say something analogous. The mental-specialists problem arises only if at least one other mental property is also totalitarian: that's what leads to the consequence that the bearers of different mental properties must have different parts. But what if none is? What if a being can remember, or think about philosophy, or recognize shapes, despite having parts *not* directly involved in those abilities? Then all these properties might be compatible, as they so obviously seem to be. We could accept the apparent fact a conscious being can also remember, recognize shapes, and so on, even though these abilities involve different parts of the brain. So maybe consciousness is totalitarian but other mental properties are not. Or perhaps some other mental property has this unique status. But no two mental properties are totalitarian.

I don't know of anyone who has actually made this proposal. It's an intriguing metaphysical conjecture. Apart from its utility in defending the two-person view, however, I can't see any reason to accept it. By itself it has no attraction. Without some account of *why* consciousness but no other mental property is totalitarian, it looks completely unprincipled. If a favourite view of mine led to a doubtful claim like this—a claim that I had no other evidence for—I'd worry. It would seriously undermine my confidence in it—especially if there were a sensible alternative. And there is a sensible alternative to all this: that conscious beings are animals.

11. More alternatives

I haven't considered all accounts of what we might be that are consistent with the two-person view. I've said nothing, for example, about the view that a person is an immaterial substance or a bundle of mental states and events. I can't see that partial twinning creates any *new* problems for these views. The view that there are two immaterial thinking beings in a case of partial twinning does not introduce any problems in addition to those arising from the view that there is one immaterial thinker attached to each ordinary human being. Someone might be led by the two-person view to conclude that conscious beings must be immaterial—not merely a dualism of mind and life, but an old-fashioned dualism of mind and body. In any event, I haven't yet seen a credible two-person view that's compatible with our being material things.[13]

13 I am grateful to Tim Campbell and Karsten Witt for comments on earlier versions.

References

Baker, L. R. (2000). *Persons and bodies*. Cambridge University Press.

Campbell, T., & McMahan, J. (2016). "Animalism and the varieties of conjoined twinning." In Blatti, S., & Snowdon, P. (eds.), *Animalism: New essays on persons, animals, and identity.* Oxford University Press.

Gunnarsson, L. (2010). *Philosophy of personal identity and multiple personality*. New York: Routledge.

Hudson, H. (2001). *A materialist metaphysics of the human person*. Cornell University Press.

Hudson, H. (2007). "I am not an animal!" In van Inwagen, P., & Zimmerman, D. (eds.), *Persons: Human and divine*. Oxford University Press.

Johnston, M. (2007). "'Human beings' revisited: My body is not an animal. In Zimmerman, D. (ed.), *Oxford studies in metaphysics* 3, pp. 33–74. Oxford University Press.

Locke, J. (1975). *An essay concerning human understanding*. Oxford University Press. (Original work 1694.)

McMahan, J. (2002). *The ethics of killing*. Oxford University Press.

Olson, E. (2007). *What are we?* Oxford University Press.

Olson, E. (2014). "The metaphysical implications of conjoined twinning." *Southern Journal of Philosophy* 52, Spindel Supplement, pp. 22–40.

Olson, E. (2016). "What does it mean to say that we are animals?" *Journal of Consciousness Studies* 22 (11–12), pp. 84–107.

Olson, E. (2018). "The zombies among us." *Noûs* 52, pp. 216–226.

Parfit, D. (2012). "Why we are not human beings." *Philosophy* 87, pp. 5–28.

Shoemaker, S. (1996). "Unity of consciousness and consciousness of unity." In *The first-person perspective and other essays,* pp. 176–98. Cambridge University Press.

Shoemaker, S. (2008). "Persons, animals, and identity." *Synthèse* 163, pp. 313–24.

Sider, T. (2007). "Parthood." *Philosophical Review* 116, pp. 51–91.

Tye, M. (2003). *Consciousness and persons*. MIT Press.

van Inwagen, P. (1990). *Material beings*. Cornell University Press.

van Inwagen, P. (1994). "Composition as identity." *Philosophical Perspectives* 8: *Logic and Language*, pp. 207–220.

John Perry
Stanford University
johnperry43@gmail.com

Original Scientific Paper
UDC 165.242.1

ON KNOWING WHO I AM

Keywords: *Castañeda, amnesia, efficiency, Kaplan, context, character, basic actions, self-informative, self-effecting, other-informative, primitive self-knowledge, self-notion, self-attached, essential indexicals.*

(August 5, 2022)

1. On Knowing Who One is. Most of us would say that we know who we are. But there are exceptions. In my version of Hector-Neri Castañeda's (1968) wonderful example a soldier – I'll call him 'Elwood Fritchey' – is involved in a battle, and performs some heroic deeds.[1] Late in the battle he is injured in a way that causes amnesia of an appropriately philosophical sort; memories of his past are inaccessible to him, but remain deep in his brain.[2] He also loses his dog tags and wanders far from the battle. He is clearly a GI, so he ends up in a military hospital. But no one can figure out who he is. The heroic soldier had blue eyes, as does Elwood, but due to an error, the records said that Elwood Fritchey had brown eyes. So the possibility that the patient was the missing hero Elwood Fritchey was mistakenly eliminated.

Consider Elwood's first few hours in the hospital. He can't remember his name or where he is from or anything else of that sort. He is given a list all of the soldiers in the recent battle, with a few key facts about each, but he can't say which one of them he is. He says, "I'm sorry, I don't know who I am".

But in an important way Elwood *does* know who he is. The nurse comes in with a tray of food. Elwood says, "Good, I am hungry." He has feelings of hunger. He knows who is hungry. He knows into whose mouth he needs to put the food, in order to relieve the hunger he feels. He knows whose hands

1 Castañeda's hero is Quintus. The sequence of events Castañeda presents is a bit puzzling. Consider, for instance, the case of a man, to be called „Quintus," who is brought unconscious to a military tent, but on gaining consciousness suffers from amnesia, and during the next months becomes a war hero and gets lost in combat and completely forgets the military chapter of his life. (p. 446)

I suspect he had the sequence of events I attribute to Elwood in mind, where the heroism comes before the amnesia, but I may be missing something.

2 See Perry (1993).

 DOI: 10.5937/BPA2336025P

and arms will move, when he decides to do so. He knows that the person who is speaking when he says, "I am hungry" is the same person whose hunger will be relieved. So, in a way, he does know who he is, in spite of not knowing his name and history before wandering away from the battle.

2. Efficiency. Ellen, the soldier in the next room, will know things about *herself* in the same way that Elwood knows about himself. If she has feelings of hunger, she will know that she is hungry, and she will know whose mouth she needs to put food in to relieve that hunger. Perceptual knowledge is *efficient*, in the sense Jon Barwise and I gave the word in *Situations and Attitudes*.[3] Different people, in different circumstances, can know or believe different things, by being in the same perceptual state. Our actions are also efficient. Different people, in different circumstances, can bring about different results, by performing the same basic actions, that is, basically moving their bodies and limbs in the same way.

Elwood and Ellen are both sitting at a bar near the hospital. A mug of beer is placed in front of each of them. They are in the same perceptual state, more or less, the state one is in when one sees a mug of beer in front of one. Elwood believes there is a mug of beer in front of him; Ellen believes there is a mug of beer in front of her: efficient perception. Elwood moves his arm and hand, picking up the glass and bringing it to his lips, opens his mouth, tilts the mug a little more, and swallows. He brings it about that he has a drink of beer. Ellen makes basically the same movements, bringing it about that she has a drink of beer: efficient action.

It's clear that Mother Nature appreciated efficiency. A species of animal has a repertoire of methods for picking up information, and a repertoire of basic actions they can execute. The billions of members of a given species will pick up different information, about themselves and the context they are in, and perform different actions with different results, on themselves and the things in a position to be effected by them.

3. Context and Character. We can borrow terminology from David Kaplan's theory of indexicals and demonstratives to elaborate on this.

An agent, location, time, and circumstances, such that the agent is in the location at the time, and the circumstances obtain in that location at that time, is a *proper context*. The *character* of an indexical expression is a function from contexts to suitable *contents*. The character of an indexical expression takes us from a context to the referent of the expression in the context. The character of 'now' takes us to the time of the context, of 'I' to the agent, of 'here' to the location. The character of a sentence with an indexical or indexicals is a function from contexts to singular propositons that includes the referents of the indexicals. So "I am hungry now" in the context consisting of Elwood, his

3 Barwise & Perry (1983).

room *r*, and the *t* moment the nurse enters, yields the singular proposition that Elwood is hungry at *t*.

Sentences with indexicals provide ways for different agents and agents in different contexts to say different things in the same way. In a debate about action I might say to Michael Bratman, probably falsely, "I am right and you are wrong". He might reply, probably truly, with the same sentence. We use the same sentence to disagree with one another.

David Kaplan's system of context and character provides a good account of the efficiency of language with indexicals. His account can be adapted for other kinds of efficiency. Philosophers of action call the movements we can normally make at will in any circumstances *basic actions*. The same basic actions, made by different agents and/or in different contexts, have different results. You mow your lawn with basically the same sequence of basic actions I do. But I mow my lawn, you mow your lawn. Different agents, different contexts, the same sequence of actions (more or less), different lawns mowed. The basic actions can be thought of having an important character-like property, that gets at a key element of their causal role. An agent who *knows how* to do things basically knows which basic actions will bring about those things in different contexts.

Elwood knows how to walk and eat and say that he is hungry using 'I'. But he doesn't know how to say that he is hungry using his name. When I wake up in the morning, look out the window, and see that it is sunny, I know how to say so using the indexical 'Today': "Today is sunny". But until I look at a calendar I won't know how to say this by using the date of the day of my waking and looking.

4. Self-Informative perceiving and self-effecting acting. Elwood has a number of ways of knowing things about himself and the things around him. There are his five external senses, by which he can know that various sorts of things are happening around *him*. For example, he knows by means of vision that the nurse he sees has tray of food. We think of the external senses as ways of finding out about other things. But they also provide information about the perceiver; Elwood knows that there is a tray of food in front of *him* – using italics as short for "in a way he would express with the first-person".

But we also have internal senses. Elwood knew that *he* was hungry through interoception, which along with proprioception, provides us with *normally self-informative* ways of knowing about what is going on inside of us and how our limbs are arranged. I say "normally" because a philosopher can imagine exceptional cases. A mad scientist sitting behind us at a movie manages to connect nerves coming from your stomach to my brain and vice versa. The more I eat, the fuller you feel. Definitely not normal. There are normally self-effecting ways of acting, as illustrated by Elwood and

Ellen's beer-drinking. But who knows that could happen in an AI lab, or a philosopher's imagination.

Introspection is a self-informative way of knowing what's happening in our own minds. Do we need to say "normally"? The mad scientists in my imagination haven't figured out how to wire us up so that I know what you are thinking by introspection and vice-versa, but give them time.

I'll say that seeing, hearing, and our other external sense provide *normally other-informative* ways of knowing about things around us, which serve as the basis for inferences about more distant things. But they are also *normally self-informative*, since one learns that one stands in various relations to the perceived objects or inferred objects. So Elwood learns in a normally other-informative way, when he sees the nurse, that the nurse *he* sees has a tray of food. He learns, in a normally self-informative way, the there is someone in front of *him* with a tray of food. He infers that there is someone who cooked the food and gave it to the nurse.

Normally other-informative ways of knowing can also be self-informative in a different way in the right circumstances. I can look at a person, read about a person, google a person, and so on. Most of the time one uses these methods to find out about other people; that's their normal use. But the same methods can be used to find out about oneself, without a philosopher or mad scientist playing a role. If I don't remember where my class meets, I'll go the appropriate university website and enter my name, and find out where *my* class meets in the same way I would find out where Michael Bratman's class meets. If I can't remember my phone number I can look it up in the phonebook – if I, unlike Elwood, know my name.

To use normally other-informative methods in this way, one needs to know how the sources of such information identify him – by his name, or library card number, or social security number, or whatever. The university library may post a list of users with the worst record of returning books on time, listed by their library card numbers. I may be amazed at the number of times the holder of #1234567 has been late in returning books, without realizing that it is me, at least until the Stanford Library Police show up at my house.

This is Elwood's situation as he looks at the information under "Elwood Fritchey" on the list of missing soliders: he is learning facts about himself without realizing that he is doing so. He learns that Elwood Fritchey was born in Broken Bow, Nebraska. He is Elwood Fritchey, but he doesn't realize that he is getting information about himself in this normally other-informative way, and has no idea where *he* was born.

A normally self-directed action is a sequence of basic actions normally used to have an effect on the agent. There is a way of scratching one's back that we each use to scratch our own backs. Such actions are often precipitated by self-informative perception, as when one has an itchy back.

I use the phrase "primitive self-knowledge" for what one knows via normally self-informative methods, and what one knows how to do using normally self-effecting actions. It doesn't include what one learns about oneself through normally other-informative methods. If my wife Frenchie tells me that I have a bit of lettuce caught in my teeth, that's not part of my primitive self-knowledge. If I learn this by feeling the lettuce with my tongue, it is.

I use the concept of primitive self-knowledge in a very broad way, to get at what I see as a key aspect of evolution, common to animals from humans to snails and beyond. Any species of living things is equipped with a range of states that, relative to constraints and circumstances, carry information about crucial things occurring in its body and environment. Being in these states cause behavior that, given that information, contributes to survival, that is, behavior that increases the probability of the living thing surviving, and/or its genome surviving, and/or the species surviving. The key point is that Nature is *efficient*. A creature's being in a state at a time indicates something about *that* creature's environment at *that* time and this state will lead to behavior useful for *that* creature (and/or its genes, and/or its species). A different member of the same species may be in the same state. It's being in that state indicates something about that creature's environment. It may lead to the same behavior. Chicken-1 sees a kernel of corn in front of her, walks forward and pecks, with the result that chicken-1 gets some nourishment. Chicken-2 see a kernel of corn in front of her, walks forward and pecks, with the result that chicken-2 gets nourishment. Mother Nature or God or whoever or whatever got our world going, realized the importance of efficiency.

Animals with primitive self-knowledge typically need to integrate information from different senses, or primitive precursors of senses, and remember information at least for a short time, and act in light of this integrated and remembered information. I call this collection of integrated information a *primitive self-notion*, when it contains only information acquired in normally self-informative ways. The primitive self-notion is not an idea that is common to all the bits of information, but more like a mental file folder into which all the bits are put. Elwood's amnesia doesn't prevent him from remembering the things that have happened to him since he wandered off from the battle, so Elwood has a primitive self-notion that is filling up with things he has observed and inferred since arriving at the hospital: *I was fed yesterday; nurses bring me food; they will probably continue to bring me food, etc.*

Most of us humans also have a lot of what I call "self-attached knowledge", which is what I take to be part of what we usually have in mind when we use the phrase "self-knowledge." If we know our name, or our social security number, or the number on our library card we can pick up information about ourselves in normally other-informative ways, the same ways we might

use to pick up information about others, and integrate with our primitive self-knowledge. Before the battle Elwood might have looked at a list on the Barracks Bulletin Board each morning to see whether he had KP duty or some other assignment that day. If he saw that Elwood Fritchey has KP duty and he felt hungry, he would know that he was both hungry and had KP duty; both will be part of his self-knowledge.

Even though non-human animals tend not to read or google or look for their names on lists on bulletin boards, they can learn about themselves in ways that are normally other-informative. A mammal or a bird can see its reflection in a pond or a mirror. In general, these bits of information will not be incorporated into the self-notion so as to interact with primitive self-knowledge and effect normally self-effecting actions.

But some animals – chimpanzees, various birds, and others – can integrate this information from a mirror with their primitive self-knowledge so their actions are motivated by by the combination. This seems to be what is involved in passing Gallup's mirror test.[4] This integrated collection is not a primitive self-notion, but something more sophisticated, which I just call a "self-notion".

If we know English, and have primitive self-knowledge, and are not worried about being trapped in the lab of a mean psychologist, we can use 'I' with confidence to express what we learn in that way. If I feel pangs of hunger I can confidently say "I'm hungry'. If I know by introspection that there is some thinking going on –-– worrying about whether I exist, perhaps –-– I know that *I* am thinking, and so, it seems, that *I* exist. Elwood can't say much of anything with confidence about where he came from or how he has spent his life. But if he limits himself to his primitive self-knowledge, he can be pretty confident that what he says his true. When he tells the nurse, "I am hungry" the nurse won't say "How do you know? You don't even know who you are!" But he eventually does learn about his past, and as his self-knowledge grows he can use 'I' to express it.

5. "Essential indexicals." The connection between self-knowledge and 'I' and other indexicals is easily overstated. 'I' is an expression of natural language, the main purposes of which is communication and retention of information. 'I' provides a way of conveying information about ourselves to people who in a position to see who is speaking. As a result, it is the standard way of referring to ourselves when people can see that we are speaking, even if they do know our name.[5] I argued in "The Problem of the Essential Indexical" that identifying ourselves (or our location, or the time of our utterance) indexically is often *essential* to explaining something to the person to whom we are speaking. I visit a doctor; she asks, "Why are you here today?" "John

4 See Baker, 39n

5 See Korta & Perry (2011), section 3.5 for a discussion of such cases.

Perry is having a lot of headaches" isn't a helpful answer, but "I am having a lot of headaches" is a pretty good explanation. Similarly, if the doctor had asked, "Why are you at 795 El Camino Real on the 6th of April" it would have just confused me, but "Why are you here today" worked just fine.

I did not mean, and did not say, that whenever a thought or action has a *character* or something like it, there must be an indexical involved, so that our brains are full of some kind of mental indexicals. I wasn't claiming that indexicals were essential to thought and action. My point was that the contribution that indexicals make to explanations, as the one I gave in the last paragraph, is hard to explain on the traditional theory of propositional attitudes. In the example "John Perry is having a lot of headaches" and "I am having a lot of headaches" express the same singular proposition. What differs is not what is said, but how it is said. We need to make the same distinction between what and how with belief states and types of actions, and so the context/character distinction is helpful. But the difference in ways of believing and ways of acting isn't typically a matter of indexicals.

If a new-born sees a nipple in front of it, it has primitive self-knowledge that will cause primitive self-directed action that fits the situation: sucking on that nipple. Other infants will do the same: efficiency. The perceptual state is efficient, the action is efficient, and the causal connection between them is to be explained in terms of general facts about infants, nipples, and nutrition. But its not efficient because indexicals are involved.

Finally, suppose you are having dinner with President De Gaulle and some others. After his first bit of steak, De Gaulle says, "Please pass the salt to President de Gaulle". Everyone knows who de Gaulle is, but his assumption that this is so seems a bit self-important. If he has just said "Please pass me the salt" it wouldn't have had this implication. Same proposition expressed, but different natural inferences made by the listeners. (CP, pp. 35–36; 83–85.)

6. On Knowing Who One is (continued). On my view, we all know who we are in basically the same way that Elwood does his first day in the hospital. Knowing who I am is a matter of knowing who I am relative to a population of identified objects, typically agents. In his room, Elwood can identify two humans, himself and the nurse. Perhaps later, at dinner, there will be dozens he can identify perceptually. If he is in doubt who he is, he can follow Wittgenstein's suggestion, raise his arm, and look around to see whose arm goes up. But if you give him a list of names, say of the soldier in the battle, he won't be able to say which one he is.

The difference between Elwood and I is that there are many more sets of identified individuals relative to which I know which one I am. The Palo Alto phonebook, for example, lists thousands of individuals; I know I am the one identified as "John Perry"; if there are more than one of them, I am the one who lives on Hilbar Lane. But given a list of emeritus Stanford faculty, identified by their university ID numbers, I no idea which one I am.

8. On Who We Are. I think all of us know who we are in basically the same way that Elwood does; we are the person about whom we have primitive self-knowledge. We are the person who is having the experiences that we are having, who is having the internal events they disclose the occurrence of, and who is perceiving external things around her via external senses: primitive self-knowledge. As we learn more about these things, including their relation to other things inferred rather than perceived, we learn more about ourselves, and develop a richer and richer self-notions, all ultimately grounded on self-informative ways of perceiving.

But this leaves some interesting questions open, questions that may not have occurred to Mother Nature but bother philosophers. Do we have essences, combinations of ordinary properties that are necessary and sufficient for being who we are? If not, do we have haecceities, non-ordinary properties, with no empirical implications, necessary and sufficient for being who we are? If neither of these things, what makes it the case that the same person does different things in different possible worlds – the problem of "trans-world identity". I'm thinking about these things. But, more and more frequently, I find myself asking, "If Mother Nature didn't have to worry about these things, do I have to?"

References

Almog, J., Perry, J., & Wettstein, H. (1989). *Themes from Kaplan*. New York: Oxford University Press.

Barwise, J., & Perry, J. (1983). *Situations and attitude*. Cambridge, Mass.: The MIT Press. (Reprinted with a new introduction by CSLI Publications, 1999.)

Castañeda, H. N. (1968). "On the logic of attributions of self-knowledge to others." *The Journal of Philosophy* 65, 439–456.

Kaplan, D. (1979). "On the logic of demonstratives." *The Journal of Philosophical Logic* 8, 81–98.

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

Korta, K., & Perry, J. (2011). *Critical pragmatics*. Cambridge: Cambridge University Press.

Perry, J. (1993). *The problem of the essential indexical and other essays*. New York: Oxford University Press. (Enlarged edition, Stanford: CSLI Publications, 2000.)

Perry, J. (2002). *Identity, personal identity and the self*. Indianapolis: Hackett Publishing.

Perry, J. (2019). *Re-visiting the essential indexical*. Stanford: CSLI Publications.

Richard Swinburne
University of Oxford
richard.swinburne@oriel.ox.ac.uk

Original Scientific Paper
UDC 165.12 Декарт Р.
14 Декарт Р.

141.112

A CARTESIAN ARGUMENT FOR SUBSTANCE DUALISM

I

In my recent book *Are we bodies or souls?*[1] I argued in favour of Descartes's version of substance dualism – that humans on earth consist of two parts, body and soul, of which the soul is the one essential part. I understand by a 'soul' a non-physical substance which is capable of having conscious experiences. We exist if and only if our souls exist. In that book I gave arguments of two kinds for this view. The first kind of argument proceeded from the impossibility of giving criteria for personal identity over time (that is, criteria for a person P2 at a time T2 being the same person as a person P1 at a time T1) in terms of the continuing of any physical substance or physical or mental property; and so the only possible necessary criterion for such identity is that P2 is the same person as P1 only if P2 has P1's soul. I shall not discuss this kind of argument in this paper. In this paper I shall summarize my treatment of the second kind of argument, an argument from the mere conceivability by each of us of our existing without a body to the conclusion that each of us has a soul as our one essential part. I shall claim that Descartes's own argument of this kind shows that the existence of the soul of each of us is a sufficient condition for our existence, but not that it is a necessary condition for our existence. I shall go on to claim that a slightly amended version of Descartes's argument will show that the existence of the soul of each of us is a necessary condition for our existence. I shall then defend this view against the normal fashionable objection to any version of any argument of this kind.

Here is Descartes's argument, as presented in *Discours de la Méthode*, which will be very familiar to you all:

> Examining attentively that which I was, I saw that I could conceive that I had no body, and that there was no world nor place where I might be; but yet that I could not for all that conceive that I was not. On the contrary, I saw from the very fact that I thought of doubting the truth of other things, it very evidently and certainly followed that I was; on the other hand if I had only ceased from thinking, even if all

1 Oxford University Press, 2019, revised edition, 2023.

 DOI: 10.5937/BPA2336033S

> the rest of what I had ever imagined had really existed, I should have no reason for thinking that I had existed. From that I knew that I was a substance the whole essence or nature of which is to think, and that for its existence there is no need of any place, nor does it depend on any material thing; so that this 'me', that is to say, the soul by which I am what I am, is entirely distinct from body, and is even more easy to know than is the latter; and even if body were not, the soul would not cease to be what it is.[2]

The argument seems to me to have three premises:

> **first premise**: I am a substance which is thinking.
>
> **second premise:** it is conceivable that 'I am thinking and I have no body'.
>
> **third premise**: it is not conceivable that 'I am thinking and I do not exist}'.
>
> from which he argues to the conclusion
>
> **conclusion** I am a soul, a substance, the essence of which is to think.

Descartes assumes that he is a 'substance'. Although Descartes gives somewhat different definitions of 'substance' in different places in his writings,[3] all that we need to assume that he means here by a 'substance' is 'a component of the world'– one of the things that exist at a particular moment of time and has properties. **Descartes's first premise** is the contingent premise that (at the time when he was considering this argument) 'I am thinking'. Descartes uses 'thinking' in a wide sense. He wrote elsewhere:

> By the word thought I understand all that of which we are conscious as operating in us. And that is why not only understanding, willing, imagining but also feeling, are here the same thing as thought.[4]

Descartes's second premise is that, while he is thinking, he can, compatibly with his logically contingent premise 'I am thinking', 'conceive that [I] have no body'. For maybe Descartes just dreams that he has a body. I shall understand a proposition being 'conceivable' as it being logically possible, in the sense of not entailing a contradiction. The premise that this is conceivable certainly

2 *The philosophical works of Descartes*, translated by E.S.Haldane and G.R.T Ross, Cambridge University Press, vol.1, 1968, p.101.

3 In some places he defines a 'substance' as a thing in which properties inhere– see, for example, his *Arguments demonstrating the existence of God and the distinction between soul and body* Definition 5, in Haldane and Ross, vol.2, p. 53. In other places he defines a 'substance' as a thing 'which can exist by itself, without the aid of any other substance' (*Reply to objections* IV in Haldane and Ross, vol.2, p.101.) Elsewhere he qualifies this second definition by acknowledging that 'creative substances... are things which need only the occurrence of God in order to exist' *(Principles of Philosophy*, principle 52, in Haldane and Ross, vol.1, p.240)

4 *Principles of Philosophy*, principle 9, in Haldane and Ross, vol.1, p.222.

seems to those who first hear it, immensely plausible. For what is it to have a body? It is to have a physical substance, a chunk of matter through which one can make a difference to the physical world and through which one learns about the physical world. Some reports of 'near death' experiences of patients undergoing an operation report that patients claim to have experiences of floating above the operating table at the same time as the surgeons certify that those patients are 'brain-dead'. And while we may suspect that really the patients did not have those experiences at exactly the same time as they were brain-dead, or that surgeons may sometimes judge a patient to be 'brain-dead' while there is still some activity in the patient's brain, we can certainly understand what the reports claim, and fairly evidently they do not entail any contradiction. And what the reports claim is that at the time of the 'near-death' experiences, the patient could not control any body or learn about the world through any body; and so their conceivable (though possibly false) claims were claims that they were having experiences at that time when they did not have a body. So– I suggest– Descartes was right in claiming that there is no contradiction entailed by 'I am thinking and I have no body', and so it is conceivable (= is logically possible) that I am thinking and I have no body'. So his second premise is true.

Descartes's third premise is that he cannot conceive that 'I am thinking and I do not exist'. And it surely is obvious that that proposition entails a contradiction and so is inconceivable. 'I am thinking' obviously entails 'I exist'; and so 'I am thinking and I do not exist' entails 'I exist and I do not exist', which is a contradiction.

So it follows from the three premises that 'I am a substance which, it is conceivable, can exist without a body'; in his own words 'for my existence there is no need of any place nor does it depend on any material thing'. Yet, if he exists, there must be some part of him which exists; and so if he can exist without a physical (= material) part, he would need a non-physical part in order to exist. And so having a soul would be sufficient for his existence. But it does not follow that he needs a soul now when he has a body, in order to exist now; and so it does not show that having a soul is a necessary condition for Descartes's existence.

However, that will follow if we substitute for what I have construed as Descartes's second premise this **amended second premise**, **that it is conceivable that 'While I am thinking, my body is suddenly destroyed',** (that is, in the middle of a period while I am now thinking my body is suddenly destroyed). I suggest that this stronger principle is also correct. If it is conceivable that while I am now thinking, I have no body, it is surely conceivable that while I am now thinking, I suddenly cease to have any control over my body or to be influenced by anything that happens in it; and in that case I would have ceased to 'have' a body; the body would be un-owned by me. Many of the reports of 'near-death' experiences of patients were of experiences of leaving their bodies; and while again we may reasonably doubt

whether the patients really left their bodies, we can understand what their claims to have left their bodies mean –that they were at one time embodied and then left their bodies and observed them from a distance. Fairly evidently these claims do not entail contradictions. So if it is conceivable that I 'observe' while I cease to have a body, it is conceivable that I remain conscious when my body has been destroyed. So 'While I am thinking, my body is suddenly destroyed' is conceivable. Then there follows from the amended second premise and the third premise this lemma that **'I** am a substance which, it is conceivable, can continue to exist while my body is suddenly destroyed'**.**

I now add what I will call **a fourth premise, that is inconceivable is that any substance can lose all its parts simultaneously and yet continue to exist**. A table may continue to exist if it loses a leg, but not if it loses all its legs and the table top at the same time. And organisms can continue to exist if over time they lose all their parts, so long as those parts are gradually replaced, each part being replaced over a period of time while other parts continue to exist. A tree can continue to exist if each cell is replaced by a similar cell at different times. But what is inconceivable is that every single part of the tree should be suddenly destroyed and yet that tree should continue to exist. It therefore follows from that, that if it is conceivable that now while Descartes has a body, that body is suddenly destroyed and yet he continues to exist, then he must actually have now also another part which is not destroyed and which is doing the thinking, and which he and I are calling his 'soul'. For if he didn't already have that other part when his body is destroyed, he could not continue to exist. And since at every time while he is thinking and has a body, it is conceivable that he should lose his body and yet continue to exist, it follows that at every time while he is thinking, he has a soul. For if at some time he did not have a soul, it would be logically impossible that at that time he should continue to exist when his body is suddenly destroyed. Hence, Descartes, knowing the truth of his first and contingent premise that he is thinking, is entitled to conclude that having a soul is not merely sufficient, but necessary for his existence. And since each of us humans who are conscious can formulate the same argument with respect to ourselves, we are entitled to conclude that each of us has a soul which is necessary and sufficient for our existence. I believe that this amended argument is valid, and – for each of us for whom its premise is true – knowably sound.

However, that puts me in a very small minority of contemporary philosophers. Contemporary philosophers might be prepared to admit that Descartes's argument is a valid argument (that is, its conclusion is indeed entailed by its premises), and they might well be prepared to admit that my own amended version is also a valid argument. But they claim that Descartes has no justification for asserting his second premise, and likewise they would claim that I have no justification for asserting the amended version of that premise. This is because, in the objectors' view, we simply have no idea of

what we are referring to by 'I'. Maybe 'I' refers to my body, and in that case of course the second premise is false. Or maybe it refers to some hidden essence of me, about the nature of which neither I nor they have the slightest idea; and in that case, although the second premise might be true, we are not in a position to know that it is true. Shoemaker has claimed more generally that that one could never reach a conclusion about the actual world from mere considerations of what is conceivable: 'it is quite hopeless to suppose that a claim of *de re* possibility, a claim to the effect that some actually existing thing could undergo such and such changes, can be grounded on mere thought experiments, or on considerations of what can be supposed or imagined without logical or conceptual incoherence'.[5] It was in order to deal with this objection that I invented some new terminology.

II

I understand by a 'designator' a word (or longer expression) which refers to some object – substance, property, event, or time. I define an 'informative designator' as a designator which is such that if we know what the designator means, we know to which object it is referring; and an 'uninformative designator' as a designator which is such that knowing what the designator means does not ensure that we know to which object it is referring.

So, first, what is it to know what some designator means? Often we know this if we know some formal definition of it, of a kind which one finds in a dictionary. A 'taxidermist' is 'a person who stuffs the skins of dead animals, in order to make them look like living animals.' A 'tort' is 'an infringement of a legal right which leads to a legal liability'. And most scientific terms have precise dictionary definitions. An 'electron' is a subatomic particle with a negative electric charge of 1.602 x10–1^{9} coulomb and a mass of 9.109 x 10–3^{1} kg. But definitions are of no use unless we know what the words in the definition mean. Those words may be defined by other words – a 'subatomic particle' is 'one of the many kinds of particles of which atoms are composed'. And so on. But in the end if we are to understand any definition its constituent designators have to be understood in some other way than by definition and that is by having an ability to recognise straight off whether they apply to an object or not. This ability is normally acquired by 'acquaintance', that is by seeing (hearing, or perceiving in some other way) objects to which those words apply and having these objects contrasted with other objects to which those words do not apply, and being told that (or coming to understand as a result of listening to conversations that) the word applies to that particular object or – alternatively – to objects like it in a certain respect. We learn what an 'animal' is, by seeing many animals, and being told that 'animals'

5 'Sydney Shoemaker's Reply', in S. Shoemaker & R. Swinburne, *Personal Identity*, Basil Blackwell, 1984, p.144.

are things like this, to be contrasted with plants which are not is like this; we learn what 'mass' means by being told that it is roughly the same as 'weight' and by being shown objects which have different weights and among them objects of 1 kg weight. We learn what '$10–3^{1}$' means by being shown what it means to reduce a quantity to one 10^{th} of its value, and what it means to perform an operation 31 times. It is in this way that we learn words (or longer expressions) designating simple observable properties or kinds of substance like 'line', 'angle', 'heavier', 'longer', 'door', 'road' 'straight', 'shirt', 'walks'; and words important in human interaction such as 'face', 'mouth', 'talks', and 'kiss'. All of us learn the meanings of a vast number of words (or longer expressions) by 'acquaintance'. For some of these words, some people learn their meaning by acquaintance, while other people learn their meaning by definition; but that can only happen if the latter learn the meaning of the words in the definition (or words by which those words are defined) by acquaintance, these latter words being words which others might have learnt by definition. In the case of words which are names of particular substances, some people learn what they mean by 'acquaintance'. 'London' is the big city in which we live; 'Stonehenge', is that arrangement of big stones in front of us; 'Woodstock road' is the road on which we live; 'the Mona Lisa' is this painting we are looking at. Other people learn that these words mean what those acquainted with the substance mean by it. This picture of the meaning of proper names is of course similar to Kripke's picture of words getting their meaning by an initial baptism, and other later speakers intending to mean by the word what the earlier speakers meant by them.[6] Clearly language is more complicated than I have been describing it, the meanings of some words depend both on acquaintance and definition, and many words are used in slightly different ways by different speakers. But with amplifications and qualifications it must be basically right. Unless the meanings of many words were determined by the objects to which most speakers would paradigmically apply them, there could be no language.

Being able to recognise instances of the correct application of a designator straight-off, as I am understanding this, involves being able to recognize whether or not it applies to an object—under ideal conditions. Conditions are ideal when one's faculties are working properly, one is in the best possible position (that is, best possible location relative to the object) for recognizing the property (or whatever) referred to, and one is not subject to an illusion. Thus if someone had normal sight and then became totally blind, their inability now to recognize a face doesn't show that they do not know what the word 'face' means. For now their faculties are not working properly. If someone is too far away from two rods, they may not be able to recognize whether one rod is 'longer than' another rod, but that doesn't show that they do not know what the expression 'longer than' means. For then they are not

6 S. Kripke, *Naming and Necessity*, Blackwell Publishing, 1980, pp. 96–7.

in the best possible position for recognising which rod is the longer. Further, the circumstances must not be illusory, that is such as to make a property (or whatever) look (feel, sound, or whatever) differently from the way it would look in paradigm circumstances (that is, the normal circumstances in which the meaning of its designator is explained to new speakers); or such as to make some other property look differently from the way it looks in paradigm circumstances, so that it looks like the property in question. It does not show that we do not know what the word 'cat' means if we cannot recognise a cat when it is disguised to look like a dog, or we misidentify a robot as a 'cat' if it is made to look and behave like a cat. Nor does it show that we do not know what 'Stonehenge' means if we cannot recognise it when all the standing stones are covered with domes, or we misidentify a perfect copy of Stonehenge made of cardboard as 'Stonehenge'. In these cases observers would be subject to an illusion.

In the case of words whose meaning we know straight-off and so are able to recognize under ideal conditions whether or not they apply, we know—simply in virtue of knowing the meaning of the word—what it is for the object to which they apply to be that object; we know logically necessary and sufficient conditions for something to be that object. For an object to be 'a door' just is for it to look, feel like, and behave like (e.g. open when pushed) paradigm instances of doors. For a person to be 'walking' just is for that person to be doing what we recognize as paradigm instances of persons 'walking' as doing when we observe them under ideal conditions (standing fairly close to that person in daylight, with eyes working properly, and not subject to some illusion). To be London just is to be the big city which we (that is, those of us who learn what 'London' means straight-off) recognize as 'London' under ideal conditions (walking around a big city, with eyes working properly, and not subject to an illusion by being in another city which looks exactly like London.) Hence these words whose meaning we know straight-off are all informative designators. Many words which denote properties and so kinds of substance (such as 'proton') (as opposed to words which denote individual substances which have those properties), and which can be defined by other words whose meaning we know straight-off, are also often used as informative designators; and so we can know to which object (that is, to which property) they refer, in the sense of knowing the logically necessary and sufficient conditions for being that object, merely in virtue of knowing their meaning. Which property a designator denotes is a matter of whether the property satisfies the definition which gives logically necessary and sufficient conditions for the application of the designator. To be a 'taxidermist' just is to be a person who stuffs the skins of dead animals, so as to make them look like living animals.

By contrast, an 'uninformative designator' is a word (or longer expression) which is such that if we know what the word means (that is, the meaning

which is common to its use in different contexts), that is not by itself enough to know to what it refers on a particular occasion of its use. Many 'definite descriptions'—that is, descriptions of an object which pick out that object by some property of that object, such as 'the tallest building in London'—are uninformative designators of that object. We may know the meaning of 'the tallest building in London' and so what property it designates, but knowing this is not enough to show us which building is the tallest building. To know this, we need to compare the heights of different buildings in London and discover the location, size, shape, and composition of the tallest one; and to discover this we need to do much empirical investigation. Likewise, most 'indexicals' are uninformative designators. An indexical is a word like 'he', or 'you', 'that river', or 'now', the referent of which (that is, to what they are referring) depends on the context in which it is uttered; that is, who says it, when, and where. Someone knows what 'that river' means iff they know that it refers to a river to which the speaker has just pointed or alluded, but unless they know the location of the river to which the speaker has pointed and where it is flowing from and to they will not know what that river is, in the sense of what makes the river that particular river. As I shall use the word 'essence', to know an 'essence' of the object is to know a set of logically necessary and sufficient conditions for an object to be that object.

There are other uninformative designators, the referent of which depends on some underlying essence which may be totally unknown. Obvious examples include those used by Kripke, Putnam and innumerable subsequent philosophers to illustrate the use of 'rigid designators' to designate kinds of substance, the logically necessary and sufficient conditions for being of that kind being having some possibly unknown chemical constitution. Thus, the word 'water' was used in the early nineteenth century as a designator of the actual transparent drinkable liquid prevalent in our rivers and seas, and of whatever has the same chemical essence as that liquid. But in ignorance of what that chemical essence was, people in the early nineteenth century did not know what it was to be water. They were able to use the word 'water' as a normally sufficient rigid designator because they knew a collection of properties which were normally sufficient for something to be water – being a transparent drinkable liquid prevalent in our rivers and seas. But they allowed that sometimes some transparent drinkable liquid to be found in some river or sea might not be water; and in ignorance of the necessary conditions for something to be water, they could not be sure whether something (liquid or solid) which was not transparent or drinkable or in our rivers or seas was or was not water. That collection of observable properties by which in practice they picked out something as 'water' is what philosophers have come to call the 'stereotype' of water. But, not knowing a set of logically necessary and sufficient conditions for something to be water (and so not knowing in a crucial sense, the 'essence' of water), they did not fully understand what it is to be water. Subsequent

scientists discovered that to be water is to consist of molecules, each of which consists of two atoms of hydrogen and one atom of oxygen.

I have been assuming that all competent speakers of a language understand a word in the same way. But many of the technical terms of a science which function as informative designators for experts of some science function as uninformative designators for the wider public. The wider public understands words such as 'gold' in what has been called a 'deferential' sense, that is they mean the word to refer to an object which they pick out by a stereotype and which also has whatever properties the experts have discovered to be essential to that object. So they understand by 'gold' a kind of chemical substance which is normally heavy, yellow, and malleable, and has whatever essence scientists have discovered the kind of substance which normally has these properties to have. I shall call an uninformative designator an '[uninformative] deferential designator' of an object iff one of the properties which determine its meaning and so help to determine whether or not it applies to an object is that an expert of some kind judges that it applies. It is involved in the meaning of an uninformative deferential designator that speakers 'defer' to some expert to tell them to what it refers. So while the wider public means something different by 'gold' from what scientists mean, both groups refer to the same chemical substance by the designator "gold" and know that they do. Likewise proper names of individual substances such as 'London' are uninformative deferential designators for those who have never visited London; they mean by 'London' that city which its inhabitants call 'London'; the latter are the experts to whom others defer for the meaning of 'London'.

Since most of the words by which we refer to properties and many of the words by which we refer to substances are informative designators, it is the case of most sentences that if we know what their constituent words mean we know what are the objects to which they refer, and so if such a sentence is logically possible, it is logically possible that the objects referred to could have the properties referred to. This is because if we can refer to an object by an informative designator, we know a set of logically necessary and sufficient conditions for something being that object, and so are in a position to work out what is logically possible and what is logically impossible for that object. Thus, not merely does the sentence 'Stonehenge is 100 million years old' not entail a contradiction and so is logically possible, but it is logically possible that the object referred to by 'Stonehenge' is 13 million years old; and this is because '100 million years old' and 'Stonehenge' (for those who know its meaning by acquaintance) are informative designators, and so in knowing what these expressions mean we know to which substance 'Stonehenge' is referring and to which property '100 million years old' is referring. So we can work out whether among the properties which it is logically possible for that substance to have is the property designated by '100 million years old'. But if a sentence contains an uninformative designator referring to, for example,

some substance such as 'water' (as used in the early nineteenth century), then the sentence may entail no contradiction but what it asserts about water may not be logically possible for the actual substance picked out by 'water'.

The distinctions which I have made in this section will allow us to analyse more precisely the nature of the objection to Descartes's argument and to my amended version of it, that we do not know to what the 'I' in what I have analysed as his second premise, refers. So in terms of my terminology the objection is that 'I' is an uninformative designator; and if that were so, it would be like most other indexicals in this respect. What I shall claim, in response to this objection, is that 'I' is an informative designator for each speaker who uses that expression, but an uninformative deferential designator for anyone who hears someone else use that expression; and that likewise one's own proper name, such as 'Richard Swinburne' is an informative designator for Richard Swinburne, but an uninformative deferential designator for everyone else. The person to whose authority other speakers must defer in order to know to what 'I' refers is the speaker is using the word: and the person to whose authority they must defer in order to know to whom the name of the person is referring is that person.

III

I shall approach the issue of whether 'I' is an informative designator of a substance by the connected issue of whether the designators of conscious experiences are informative or uninformative. By a person's 'conscious experiences' I mean those events about which a person knows better than anyone else can whether or not they are occurring in them. These events include sensations, occurrent thoughts, and intentions. I shall illustrate this point only with respect to sensations. Each of us has privileged access to our sensations in the sense that whatever way anyone else has of learning whether or not we now have a headache or are hearing a noise, or feeling sick, we could always use the same way as other people use to learn about whether or not we are having these experiences, yet we have a unique way which is not available to others – we know which kind of sensation we are having because we are experiencing it. The purpose of this approach is to argue that the words referring to each person's conscious experiences are informative for that person, but not for others; and so it is to be expected that the word referring to that person themselves, 'I', is informative for that person, but not for others.

We may mean by a word which we use to describe a current sensation, either simply a sensation of the kind which we are currently experiencing, or a sensation of a kind which we have experienced in the past and can recognise its reappearance. Because we have this unique way of access to our sensations, the person having a sensation is the only person who is ever in

the best possible position to know the content of that sensation, that is what kind of sensation it is in the sense of what the sensation feels like. The only faculty involved in our awareness of our sensations is our faculty of awareness, that is being conscious; and so inevitably our faculties are working properly when we have sensations. Hence – unless we are subject to an illusion (in falsely supposing the sensation to be similar to certain past sensations) we can always recognise whether some word whose meaning we know applies to that sensation or not. Hence the condition for a word whose meaning we know being an informative designator is always satisfied with respect to all the words which we use to describe our own sensations.

We normally use words to describe different kinds of sensations by the physical events which normally cause sensations of that kind and/or by the physical events (or the desires to do certain actions) which they normally cause. We learn to describe some food as 'tasting of coffee' iff it tastes like the taste which coffee causes. We learn to describe a smell as a 'smell of burnt almonds' iff it smells like the smell which burnt almonds cause. We learn to describe an after-image as a 'red' after-image iff it has the same visual appearance as do the visual appearances of (and so caused by) certain paradigm public 'red' objects (British post boxes, ripe tomatoes, strawberries etc.). We learn to describe a sensation as an itch iff it is the kind of sensation which causes us to scratch (or to desire to scratch) the place which seems to cause the sensation. For some kinds of sensations both their normal causes and their normal effects are important for enabling us to pick out the sensation to which we are referring. We learn to describe a sensation as an 'acute pain' iff it is the sort of sensation which is caused in us by certain kinds of bodily events (such as being cut or burnt), and which causes aversive behaviour (causes us to try to stop the pain if we know how to do so, for example by withdrawing a hand from the cause of the cut or running away from the fire which is burning us, or strongly to desire to do so). But having learnt to refer to a particular kind of sensation by events which normally cause it or are caused by it, we are then in a position to refer to a sensation of that particular kind when it does not have these normal causes or effects. What each of us means by a 'red sensation', 'taste of coffee', 'smell of burnt almonds' is what we experience, not the normal cause of that experience.

The fact that most other people learn to use words denoting sensations in the same way as we learn them, and seem to make the same distinctions between different sensations as we do (for example distinguishing red sensations from green sensations, and the taste of coffee from the taste of chocolate) makes it probable that they mean by the words they use to describe their sensations what we mean by the same words. But it can sometimes be reasonable for each of us to doubt whether other people do mean by the words they use to describe their sensations the same as we do. Some people cannot distinguish between red and green objects; red and green objects both

look the same to them. So either red objects do not look to them the way they look to most of us and/or green objects do not look to them the way they look to most of us; and so either when they say 'it looks red' or 'it looks green;', they must mean that it has a sensory appearance different from the sensory appearance which we describe by these sentences. (Note that whether some public surface is 'red' is a public matter, a matter of whether it looks to most observers to be of the same colour as (to produce the same colour sensations as) certain paradigm objects. But whether a surface 'looks red' to a certain person – as I am using that expression – is a matter of the particular sensory quality which it has in that person's experience of it.) And in the case of many sensations and especially tastes, the different reactions which people often have to the same input from their sense organs supports the hypothesis that the sensations caused thereby are sometimes different in different people. Insofar as we have reason to suppose that experiencing certain physical events causes different sensations in others from what it does in ourselves, we have reason to suppose that they mean something different by the words they use to describe their sensations from what we mean by those words.

It follows that expressions used by us to describe our own sensations function as informative designators for us of our own sensations, but when used by others to describe their sensations serve only as uninformative designators for us. Suppose that both John and Mary learn to use 'acute pain' as a designator of the kind of sensation which is normally caused in them by certain kinds of bodily events (such as being cut or burnt); and which normally causes aversive behaviour (for example, causes them to withdraw their hand from the cause of the cut or run away from the fire which is burning them, or to desire to do so). Then for John 'acute pain' just means the sort of sensation which is caused *in him* by those bodily events and which causes *him* to show aversive behaviour. Whether or not on a particular occasion it has such causes and effects, he uses 'acute pain' to refer to the intrinsic character of any sensation which feels like that. But he will understand Mary's use of 'acute pain' to mean the sort of sensation to which Mary alone has privileged access, which is normally caused *in her* by those bodily events and which normally causes *her* to show aversive behaviour, whether or not on a particular occasion it has such causes and effects – although he will also believe, in the absence of contrary evidence, that probably what she means is the same as what he means.

John's understands Mary's use of 'acute pain' as a 'deferential' uninformative designator because John understands Mary's use of that expression to mean whatever Mary means by whatever sensation of hers is normally picked out by the stereotype of certain bodily manifestations. The situation of John in understanding Mary's use of 'acute pain' to mean whatever she means by a sensation related to certain observable manifestations, is thus like that of the unscientific public who understand the scientists' use of the word 'gold' to mean what the scientists mean by the essence of an object which normally

has certain observable properties. But there is this crucial difference from the scientific case, that while the unscientific public can – if they so choose – learn to use 'gold' as an informative designator, John could never learn to understand Mary's use of 'acute pain' as an informative designator. This is because Mary alone can ever be in the best possible position to know whether that expression as used by her applies to some sensation of hers. And of course conversely – Mary can never learn to understand John's use of 'acute pain' as an informative designator.

IV

Conscious events are the events they are, not merely in virtue of the properties, such as 'having a pain' or 'smelling a smell of burnt almonds' involved in them, but in virtue of the persons who have these properties. So finally, are the words by which we refer to persons informative or uninformative designators? I suggest that the understanding of the criteria for some word being an informative designator developed in this paper enables us now at last to answer this crucial question. As when we analyse the meanings of words used to refer to conscious events, we need to distinguish the meanings of words used to refer to our own conscious events from the meanings of words used by others to refer to their conscious events, so too in analysing the meanings of words used to refer to persons, we need to distinguish the meanings of words used to refer to ourselves from the meanings of words used to refer to others. We refer to ourself by the word 'I'. I suggest that, as used by each person, 'I' is an informative designator of themself. For we are always able, when (1) our faculties are working properly, (2) we are in the best possible position for recognizing ourself, and (3) we are not subject to an illusion, to recognize when some person is 'I' and when it is not. When each of us refers to themself as the subject of a current conscious event, the faculty which they need for this purpose is clearly in working order. Each of us is in the best possible position for recognising themself when they pick out themself as the subject of a current conscious event, as the person who is now having this pain or that thought. Under those circumstances none of us can possibly be subject to illusion. For an illusion would consist in the circumstances being such that I refer to someone as 'I' who is not myself, or I fail to recognize myself as 'I'. But I couldn't possibly have a pain and suppose that really it was someone else who was having the pain; nor could someone else be having a pain and I suppose mistakenly that the pain is really mine. Of course, I could suppose that some pain which I am having was of just the same kind as a pain which someone else was having, but what I cannot be mistaken about is that I am having the former pain – because if I thought that I was not having a pain, I wouldn't be feeling anything. When each of us is referring to themself as the

subject of a current conscious event, we are in Shoemaker's phrase, 'immune to error through misidentification'.[7]

Since (unlike most other indexicals) 'I' as used by me is an informative designator –for me, so too is 'Richard Swinburne' – for me; and so is 'I' and their own proper name – for each other person. So the situation is similar to the situation with respect to description of our sensations. Just as each of us is always in better position than anyone else could be with respect to describing their sensations, so too each of us is always in a better position than anyone else ever could be for recognizing ourselves – and that is when we refer to ourselves as the subject of sensations (or other current conscious experiences). Others can only pick us out as that human who has that particular body and/or brain and/or makes certain memory claims (and perhaps has a certain character) and refers to himself or herself as 'I' or by their own proper name. But since those others do not have the (not merely privileged but infallible) access to who the person is to whom they are referring which I have, their use of 'Richard Swinburne' or some indexical expression to refer to me involves using it as an uninformative designator. Others mean by 'Richard Swinburne' the person whose body is such and such a body, or whose brain is such-and-such a brain, and/or who makes certain memory claims, and refers to himself as 'Richard Swinburne' or 'I'. Their use of the uninformative designator 'Richard Swinburne' is therefore also deferential; they regard me, picked out by physical properties, one of which is that I (picked out by the other physical properties) refer to myself as 'I', as the expert on who I am. But I mean by 'I' the person who is aware of himself as experiencing a certain particular conscious event, and not any person who is not experiencing that conscious event.

Our infallible knowledge of ourself is an infallible knowledge of ourself as existing at the moment at which we are aware of this. We do not have infallible awareness of what we experienced at some past time or even whether we existed at some past time, nor any infallible awareness of what we will experience in future. Nevertheless, when I believe that I experienced such-and-such at a certain past time, or will experience such-and-such at a certain future time, I know infallibly who it is to whom I believe that such and such experiences occurred or will occur. Others who believe that I experienced such-and-such at a certain past time or will experience such-and-such at a certain future time cannot know as well as I do what it would be like for their beliefs to be true because they do not have the infallible access which I have to the identity of the person about whom they have these beliefs.

It follows that Descartes did know to what he was referring by 'I' when he claimed that it was 'conceivable' (= logically possible) that 'while I am thinking, I have no body'. So, knowing a set of logically necessary and

7 Sydney Shoemaker, 'Introspection and The Self' in (ed.) Q. Cassam, *Self-Knowledge*, Oxford University Press, 1994, p.82.

sufficient conditions for a person to be 'I' (consisting in being the person who is having certain particular experiences of which he is aware), he knew what it would be like for that proposition to be true, given to what 'I' refers; he was in a position to judge whether or not that proposition is conceivable. And the same goes for the similar proposition used in my amended version of Descartes's argument, 'While I am thinking, my body is suddenly destroyed', which I claimed to be conceivable (= logically possible), given what 'I' refers to– a claim which I hope that I made plausible by spelling out one way in which it could be true. So this (now) traditional objection to Descartes's argument that no one knows to what they are referring by 'I', which – if cogent – would apply also to my revised version of it, fails. And since each of us can use Descartes's argument to show the same thing about him or herself, all human beings are substances having one and only one essential part, their soul. Each of us also has a body, but our body is a non-essential part of us. We are who we are independently of the body to which we are linked; and – if it were naturally possible for the stream of our consciousness to continue when our body ceases to function – it would be our soul and so we who continue to exist.

Živan Lazović
Faculty of Philosophy
at the University of Belgrade
zlazovic@f.bg.ac.rs

Mirjana Sokić
The Institute for Philosophy
at the University of Belgrade
mirjanasokic19@gmail.com

Original Scientific Paper
UDC 165.1
004.81:165.1

ARTIFICIAL THINKERS AND COGNITIVE ARCHITECTURE

Abstract: *This paper aims to propose and justify a framework for understanding the concept of personhood in both biological and artificial entities. The framework is based on a set of requirements that make up a suitable cognitive architecture for an entity to be considered a person, including the ability to have propositionally structured intentional states, having a form of sensory capabilities, and having a means of interacting with the environment. The case of individuals in a persistent vegetative state, as studied by Owen, serves as an example to show the importance of each of these requirements and the possibility of a "hybridization" of personhood. The proposed set of requirements provide a complete framework for understanding the concept of personhood and highlight the significance of cognitive architecture in determining personhood.*

Keywords: *Personhood, cognitive architecture, artificial intelligence, artificial thinkers, hardware, program.*

1. Introduction

Isaac Asimov's seminal work, *Bicentennial Man* (1990), presents a thought-provoking exploration of the concept of artificial personhood through the characterization of its protagonist, Andrew, an advanced robot engineered with cutting-edge technologies and designed to resemble and emulate human behavior. The novel's narrative progression, which is marked by Andrew's interactions with human beings and his subsequent questioning of his own identity, raises important queries about the definition and determining factors of personhood. Asimov's novel posits that the concept of personhood may not be reducible to physical characteristics or capabilities alone, but rather encompasses a complex interplay of consciousness, emotions, and intellect.

 DOI: 10.5937/BPA2336049L

The novel's treatment of this question leaves Andrew's personhood open to interpretation and invites readers to reflect on their own perceptions of what it means to be a person.[1] Central to the narrative is the question of whether it is the hardware or advanced software that imbues Andrew with his sense of self and humanity. Additionally, the novel raises ethical implications of this question and its potential impact on the understanding of personhood in the context of artificial intelligence. This study aims to proffer a response to the inquiry pertaining to the attributes that endow some artificial entities or systems with the capability of being deemed as artificial persons. Furthermore, it serves as the underlying foundation for a comprehensive examination of the issues pertaining to synchronic and diachronic personal identity, as well as various other relevant philosophical concerns.[2] Yet, prior to engaging in these contemplations, it is imperative to furnish a more exhaustive explanation of the philosophical framework in which this discourse is situated.

As technology advances at a rapid pace, the possibility of creating authentic thoughts and consciousness in artificial systems becomes increasingly plausible. This has led to significant attention being paid to the field of artificial intelligence within both scientific and philosophical communities, with much of the discourse centered on determining whether programming a computer in a specific way can result in the production of authentic, conscious thought. However, as Eric Olson (2019: 69) points out, these debates often overlook the fundamental aspect that a thought can only exist in the presence of a "thinker" – i.e. an entity that serves as the embodiment or manifestation of that thought. This raises the question of the nature of the artificial thinker, and prompts the inquiry into what the *subject* of these artificial thoughts would be. The most common answers to this question are that (a) the *computer* itself would be the intelligent subject or (b) that it would be the *program* running on the computer.

These answers are often accepted without further argumentation or critical evaluation, presenting a significant challenge in the field of artificial intelligence and calling for further exploration and analysis of the nature of the artificial thinker. The philosophical problem that directly arises from this uncritical response to the posed question is highlighted by Olson in the following passage:

1 It is important to note that the physical resemblance of Andrew to a human is not a crucial determinant in assessing his personhood. The purpose of this research is to establish the necessary and sufficient conditions that any biological or artificial entity or system must fulfill in order to be considered a person.

2 The distinction between diachronic and synchronic personal identity can be described as follows: diachronic personal identity refers to the continuity of an individual's identity across different stages of their life, including their memories, personality, and physical characteristics (see Maslin 2001: 242; Noonan 2019: 14). In contrast, synchronic identity refers to the characteristics and attributes that an individual possesses at a particular point in time. It can be conceptualized as a snapshot of the individual's identity at that specific moment, including all the elements that make up their identity at that time.

> For there to be thought or consciousness is for there to be *something* that thinks or is conscious – just as for there to be life is for there to be living things, and for there to be movement is for something to move. For there to be artificial intelligence, then, there must be *an artificially intelligent being: a thing that is intelligent* because of what a computer does. So there are two different questions concerning the possibility of artificial intelligence. One is whether anything in the nature of thought itself prevents it from occurring in computers. We might call this the *question of artificial thought*. The other is whether anything could be an artificial thinker. We might call this the *question of artificial thinkers*. The second question has to do with the *sort of entity an artificial thinker would be. What properties would it have, in addition to its mental properties? Would it be a material thing?* If so, what matter would make it up? If not, what sort of immaterial thing could it be? What might it be made of if not matter? (2019: 68, emphasis added)

It is important to note that Olson does not provide a comprehensive justification for his assertion that every thought necessitates a bearer, instead treating it as an axiom in his examination of the issue of artificial intelligence (2019: 69–70). In this paper, we accept the thesis that thought, whether in the context of biological (natural) or artificial systems, must have a carrier. In other words, we maintain that the presence of a thinking entity is necessary for the existence of thought. This claim is rooted in the understanding that thought emerges as a property of complex systems and therefore requires a substrate or carrier to exist and manifest. This thesis holds substantial implications in both cognitive science and philosophy, and has significant implications for our understanding of the nature of thought and consciousness in both biological and artificial systems. Additionally, we acknowledge that Olson's analysis highlights important challenges with the concept of artificial intelligence.[3] The most significant of these problems can be summarized as follows:

1. The term "artificial intelligence" typically refers to the possibility of creating streams of conscious thoughts, but these thoughts cannot exist without an artificial thinker.
2. The ontological question of artificial persons (or thinkers) is largely neglected in scientific and philosophical discussions.

---

3 Olson's use of the term "intelligence" is somewhat idiosyncratic. In his understanding, this term refers to mental phenomena such as beliefs, desires, emotions, consciousness, etc. – that is, thoughts and consciousness in general. He also notes that the term "artificial intelligence" in its common sense is often used to refer to forms of intelligent *behavior* in computers and machines (e.g. sorting different types of shapes or materials, playing chess, automated cars, etc.). In contrast to this common use, Olson refers to this term as the conceivable possibility of producing *authentic thoughts* and consciousness in artificial systems such as computers (see Olson 2019: 67).

3. Instead of a precisely specified concept of an artificial thinker, discussions about artificial intelligence often use vague terms such as "system", "substrate", or "medium", the reference of which is unclear. (Olson 2019: 69)

We concur with Olson's assessment that the field of artificial intelligence is beset by misunderstandings and ambiguities in terminology. In this research, we propose a framework for determining personhood in entities, based on the fulfillment of certain conditions such as propositional intentional states, sensory apparatus, and an apparatus for interaction with the environment. This framework, which we term "the architectural view", provides a comprehensive and unified understanding of the concept for both artificial and biological entities.

The structure of the remaining course of this paper is as follows: Initially, we will undertake a more comprehensive examination of the perspectives known as the "hardware view" and the "program view", as well as the criticisms they face. Subsequently, we will conduct a critical analysis of Olson's conceptualization of the notion of artificial persons. This critical analysis will demonstrate the need for a hybrid understanding of the notion of persons, both in biological and artificial contexts, one that circumvents the drawbacks of both aforementioned perspectives while retaining their advantages and drawing upon the concept of cognitive architecture. In the final section, we will evaluate the cogency of our hybrid proposal. Our aim is to conclude that our focus on cognitive architecture allows for a more coherent examination of the implications of personhood in various contexts and sheds light on the nature of both biological and artificial thinkers.

2. The hardware view

As advancements in technology pave the way for the possibility of simulating streams of conscious thoughts within computers, a fundamental question arises regarding the agency responsible for such thought processes. According to Olson (2019: 70), it is commonly assumed that the computer, as the physical embodiment of the system, would be the entity engaged in the act of thinking (see e.g. Turing 1950; Putnam 1964; Searle 1980: 417; Haugeland 1985; Russell & Norvig 2010).

Despite its prevalence, Olson notes that this answer is rarely supported by further arguments and is often accepted uncritically in debates about artificial intelligence (Olson 2019: 70). In addition, discussions of the nature of artificial intelligence often do not specify the type of objects that these thinking computers are: it is simply assumed that the intelligent entity is a *physical object* made of metal, plastic, and silicon chips. Olson refers to this assumption as "the hardware view" and asserts that it is the best answer to

the question of the nature of artificial thinkers. However, it is important to note that this view is not without its significant criticisms. The main obstacle to arriving at a satisfactory answer to the question of the nature of artificial thinkers in the spirit of the hardware view is the presence of the following two assumptions, which align closely with our common-sense intuitions and appear in discussions of the nature and conditions of the diachronic identity of biological persons.

One assumption is that programming a computer for intelligence does not simply *bestow* intelligence upon a previously non-intelligent being, but rather *creates* an intelligent being. This would imply that installing and uninstalling a program results in the creation and destruction of an intelligent thinker. However, it is clear that such actions do not affect the physical hardware of the computer. This leads to the conclusion that artificial thinkers and computers, which are identical in terms of hardware, have distinct histories or conditions of persistence. For example, the computer hardware would exist before and after the existence of the intelligent thinker (Olson 2019: 71).[4] In other words, the hardware view is confronted with the problem of diachronic identity in relation to artificial persons.

In the context of biological individuals, Olson presents a solution to the problem of diachronic identity that is based on the preservation of numerical identity. Specifically, he argues that an individual, *A*, at time t_1, is numerically identical to another individual, *B*, at a future time t_2, if and only if *B* at t_2 possesses the same biological organism as *A* at t_1. To elaborate, the individual who defeated Persian King Darius III at the Battle of Guagamela is numerically identical to the individual whose teacher was Aristotle years prior to that event, as the individual is the *same biological organism* belonging to the species Homo sapiens. Olson posits that psychological characteristics do not play a role in determining diachronic identity; for instance, an individual in a coma, lacking all psychological activity and content, is still considered the same person as the individual who was once a renowned Formula 1 driver, as they represent the *same living organism*. According to Olson, this illustrates that the preservation of numerical identity is dependent solely on the biological organism and not on any psychological characteristics or traits (see Olson 2000: 16–18).

It is important to note, however, that this solution is not applicable in the context of artificial persons. In the context of biological individuals belonging to the species Homo sapiens, a fetus and newborn, while lacking psychological characteristics that would classify them as individuals, still represent living organisms that will develop these characteristics through natural development. Thus, the emergence of psychological characteristics

4 It also raises the well-known problem of "too-many-thinkers" in the context of artificial intelligence. For further information regarding this problem, see Snowdon 1990; Olson 2000; Sutton 2014.

that define an individual is an inherent aspect of biological organisms belonging to the species Homo sapiens. In contrast, a collection of processors and silicon chips can exist and function without ever becoming an artificial thinker; this capability is only achieved through the pairing of appropriate software with the hardware. This shows that the reduction of artificial persons solely to hardware does not offer a convincing solution to the problem of diachronic personal identity. In simpler terms, Olson posits that the hardware view fails to furnish a convincing resolution to the issue of the diachronic identity of artificial thinkers or persons, in a manner similar to how animalism addresses the issue in the case of biological persons.

Another assumption that is often made in the context of artificial intelligence is that an intelligent entity can be transferred from one piece of hardware to another through the transfer of *data*, such as a *program file*. This would imply that the first piece of hardware would lose all of its mental properties, including memories, beliefs, preferences, and cognitive abilities, while the second piece of hardware would acquire them after the program is installed. However, this assumption is incompatible with the hardware view, for it implies that an intelligent entity possesses a property that hardware does not have, namely the ability to be transferred through data transfer; i.e. allowing for such a transfer would entail that the artificial thinker or person is identified with the program, rather than the hardware (Olson 2019: 71).

One potential solution proposed by Olson for addressing the issues associated with the hardware view is to adopt the perspective that programming a computer does not *create* an intelligent entity, but rather, it *imbues* a previously unintelligent entity with intelligence. Analogously, the deletion of data or software does not eliminate an intelligent entity, but rather renders it non-intelligent. Additionally, proponents of the hardware view may reject the idea that an artificial thinker can be transferred from one piece of computer hardware to another, despite the transfer preserving the psychological continuity of the artificial thinker. In other words, they may argue that when the hard disk of one computer (C_1) is transferred to another computer (C_2), it should be seen as C_2 receiving a new hard disk with new programs, rather than C_2 becoming C_1. This avoids the challenges associated with the requirement for psychological continuity for the diachronic identity of artificial persons. This strategy would allow the hardware view to explain the nature of artificial persons using the same (anti-psychological) model as animalism does for biological persons. However, this solution encounter counterintuitive conclusions: if we transfer a hard disk containing a program that creates a conscious artificial person, "Eve", from computer C_1 to computer C_2, it appears evident that Eve will be transferred from C_1 to C_2, which is not analogous to simply switching cables or components from C_1 to C_2. As such, it seems that not all organs or computer components are equally crucial for the realization of persons, and this will be further examined in Section 4 of the paper.

3. The program view

In this section, we will consider the so-called "program view", according to which artificial thinkers are not *physical entities*, but rather computer programs running on a computer; i.e. the concept of an artificial thinker or person is perceived as an *intangible* or *virtual* construct rather than a tangible entity.[5] The program view allows for the overcoming of two common-sense assumptions that have presented challenges for the alternative hardware view.

One of these assumptions is the thesis that the initiation and termination of a program results in the creation and destruction of an artificial person. This thesis can be further refined by considering the AI program as enabling the computer to attain conscious mental states. In this context, the initial installation of the program on a computer can be understood as the "birth" of an artificial person, while turning off and on the computer represents putting the person to sleep and waking them up; modifications to the program can be seen as modifications to the person's psychological content and abilities, and uninstalling the program or resetting the computer can be seen as the person's "death".

Another advantage of the program view is that it accounts for the transfer of an artificial person from one piece of hardware to another in a straightforward manner. In other words, it follows from this view that by transferring the data (program file) from one computer to another, the first computer loses all of its mental properties and the second computer acquires them. This implies that the artificial person possesses the property of being able to be transferred from one piece of hardware to another, which is not possible in the hardware view. One limitation of the program view of artificial intelligence is that it suggests that artificial persons may be incapable of intrinsic change. This limitation arises from the fact that the type of program that enables artificial intelligence does not change when the computer is turned on or when a sentence is typed and saved. As Olson points out:

> [U]niversals don't change. ... any more than the colour white changes when I spill coffee on a piece of paper. ... At most a particular concrete instance of the program can change. But conscious, thinking beings must be able to change intrinsically: in their beliefs, preferences, and perceptual states. (2019: 74–75)

In short, if artificial persons are identified with a certain *type* of program, they would not be able to experience changes in their beliefs, preferences, and perceptual states. However, the conclusion that the program view is ultimately

5 Olson correctly observes that, from this perspective, an artificial person would be viewed as a set of instructions, and as such, it would be brought into existence at the time of its initial conception or notation, rather than upon initiation of execution on a computer (see Olson 2019: 74)

flawed may be premature, as it is not immediately evident why proponents of this view could not claim that each artificial person is a particular concrete *instance* of the program type, rather than being identical to the program type itself. This alternative formulation of the view aligns more closely with the intuitive understanding of biological persons and allows for the possibility of changes to individual instances of the program.[6]

Despite this, we will conclude this section by stating that Olson effectively brings attention to the significant drawbacks of the program view, and that such limitations serve as a counterargument to the acceptance of a viewpoint in which the notion of a person within a biological context is fully identified with or even reduced to a set of psychological characteristics. In the following section, we will thoroughly examine Olson's ideas on artificial persons, which will enable us to articulate our own position.

4. Objections to Olson's analysis

According to Olson's view on personal identity, which is generally known as "animalism", a person is identical to a specific biological organism (Olson 2000: 16). As previously discussed, the hardware view adopts a similar approach in attempting to explain the concept of artificial persons, suggesting that an artificial person is merely a specific material object. However, Olson argues that there are fundamental distinctions in the definitions of biological and artificial persons that pose significant challenges for the hardware view. We take it that Olson's argument contains several shortcomings which we will further examine in this section.

First, the intuitive appeal and acceptance of Olson's position on the nature of biological persons can be attributed to the fact that, due to the current technical inability to transfer brains from one body to another, the concept of self has been indelibly linked to the physical body that realizes it. Even in the event that brain transfer becomes feasible, it is likely that many individuals would reject the notion that person *S*, previously embodied in body B_1, would remain the same individual post-transplantation into body B_2. It is reasonable to assume that many acquaintances of person *S* would likely not accept that they are interacting with the same person in a new body, instead positing that *S* continues to reside within the original body B_1, despite the fact that it no longer possesses the first-person perspective or psychological attributes characteristic of person *S* as they knew it. These intuitions make it challenging to accept the notion that psychological characteristics constitute an integral aspect of one's identity. It is important to note, however, that human intuitions do not necessarily serve as a reliable indicator of truth.

6 For instance, it can be argued that every human individual is a particular concrete instance of the type of biological organisms within the genus Homo sapiens.

The history of philosophy and scientific inquiry is replete with instances in which intuitive claims have been proven false, as well as those in which counterintuitive assertions have been vindicated.

The conventional and commonly held understanding of the concept of personhood extends to artificial entities as well. However, at present, the state of computer programs and artificial intelligence development does not afford us the capability to establish significant communication and interpersonal relationships with a distinct artificial person. Transferring digital files from one drive to another is not perceived as being fundamentally different from simply moving digital material from one location to another. However, as technological advancements progress, it will become possible to "transfer" an artificial person from one medium to another. The digital structure or program with which we establish communication will assume a far more significant role than the hardware components have thus far. These components will be viewed solely as the physical embodiment of an intelligent entity, rather than being identified as the entity itself. This may also lead to new intuitions regarding biological persons, who will no longer be seen as being inextricably linked to a single physical embodiment, as has been the case thus far. This makes the work of philosophers and cognitive scientists on determining the concept of artificial persons even more vital, as it can help us transcend some of the preconceptions that have been formed through familiar thought experiments based on current examples of brain transplantation or memory transfer.

A further component of Olson's argumentation with which we do not concur pertains to the assertion that there exists a clear distinction between the way in which we can explain the nature of biological persons on one hand, and artificial persons on the other. Olson articulates this thesis in the following manner:

> *An organism's life* is roughly the sum of its physiological, immune, and metabolic activities. *My* hands are parts of me because they are caught up in *my* life: they and all their parts are nourished by my bloodstream and participate in *my* metabolic processes. *My* gloves are not parts of me because they are not caught up in *my* life. There are many hard questions about what counts as *a life*, but this is at least a start. Obviously nothing like this could apply to artificial thinkers. What would be the corresponding principle for them? If my computer's central processing unit could be a part of an artificial thinker but its keyboard could not, why should this be? ... The keyboard does not seem to be involved in the computer's thought at all. And although its power supply is involved--the computer could not produce thought without it – its involvement seems only indirect, compared to certain parts of the computer's digital circuitry. This suggests that an artificial thinker would be composed entirely of electronic components and the wires connecting them. It

> would be a thin, spidery thing made of metal and silicon weighing only an ounce or two. (Olson 2019: 79, emphasis added)

It is our understanding that this passage contains several highly problematic points that are deserving of careful consideration. Primarily, Olson makes a transition from a discourse on an unspecified organism, referred to through the use of the appropriate indefinite pronoun "a", to a discourse on what constitutes an individualized life from a first-person perspective, utilizing the personal pronoun "my". It is well-established that certain physiological, immune, and metabolic processes are necessary for the maintenance of life in biological entities. This is accurately reflected in Olson's assertion that these processes serve as the foundation for the continuation of life in biological systems.

However, it is important to note that these activities are not sufficient for the life in question to be developed enough to allow for a first-person perspective, as Olson later on asserts. Specifically, immediately following the enumeration of these basic systems necessary for maintaining a biological life, Olson, without any additional justification, speaks in terms of "my" metabolic processes. While the physiological activities previously mentioned may be adequate for enabling and maintaining the life of "a" biological organism, they cannot, without the presence of additional and highly complex activities within the brain enabling a first-person perspective, enable "my" life. In summary, the necessary conditions for maintaining "a" biological life are indeed necessary, yet not sufficient for the maintenance of "my" life. It is clear that acknowledging this reality is not consonant with the radical biological viewpoint that Olson has espoused for over three decades; nonetheless, disregarding this fact and Olson's shift from the indefinite pronoun "a" to the personal pronoun "my" is entirely unjustified and lacking in explanatory rationale.

Furthermore, Olson's argument that the basic components of a computer, such as wires and simple electrical components, would be insufficient to support the operation of an artificial thinker or person, is problematic. This is because it overlooks the fact that, similar to biological entities, the functioning of these components alone does not warrant the attribution of characteristics such as a first-person perspective or intelligent thought to an artificial system. This argument is problematic because it does not consider the more complex systems and operations that are necessary for an artificial entity to be considered a thinker or a person. Such complex systems include those that allow for cognitive ability, decision making and self-awareness. These systems would be composed of not just the basic components such as wires and electrical components but many other sophisticated elements which makes it more complex to attribute the characteristics of an artificial person. In the case of biological entities, the fact that they possess basic metabolic activities does not alone suffice to attribute to them the property of a biological person. Similarly, the mere functioning of the basic components

of an artificial system does not warrant the attribution of a first-person perspective, intelligent thought, or any other characteristics that would make it an artificial thinker. Therefore, this objection, which appears to be the main objection that Olson presents to the hardware view on the nature of artificial persons, is not supported by valid reasoning and lacks foundation.

Despite these objections, we agree with Olson's assertion that for an artificial entity to be considered a person or thinker, it must have at least some sort of *material constitution*, similar to that of biological entities. This assertion goes against the prevailing psychological understanding of personhood, the implication of which is that a person can be viewed as a purely abstract entity or functional structure that is capable of being "transferred" and manifested in different physical forms, and may potentially exist without any physical manifestation (Olson 2019: 74).[7] We contend that this prevailing understanding of the necessary and sufficient conditions of personhood in both biological and artificial entities is inadequate. To address this, we put forth an alternative account of personhood, referred to as the "architectural view", according to which an entity must fulfill certain objective criteria of material constitution to be regarded as a person. These criteria will be discussed in detail in the following section of this paper.

5. The architectural view

According to the architectural view, the notion of personhood is closely linked to the presence of a specific physical structure, commonly referred to as *cognitive architecture*. The concept of cognitive architecture is rooted in the works of cognitive scientists and philosophers, such as Pylyshyn, as explored in publications such as Lepore & Pylyshyn (1999). In its simplest form, cognitive architecture can be understood as the *appropriate structure* that enables the emergence of intelligent behavior (Milojević, 2018: 183), or, alternatively, as the *underlying infrastructure* for an intelligent system (Langley et al., 2009: 141).

More specifically, in his book *Macrocognition* (2013), Bryce Huebner explains that cognitive architecture

> consists of relatively independent subsystems, which each process a narrow range of information, and which can be coordinated and

7 We wish to clarify that, although we have adopted and defended the functionalist perspective on the nature of mentality in previous works (see Lazović 2009 and Sokić 2020), we reject the implication that is often associated with functionalism in the context of personhood. This implication holds that the essential property of a person is seen in the abstract functional structure from which propositionally structured intentional states (as well as other mental states and psychological contents) can be attributed. Our understanding, as presented in this paper, is that for an entity to be considered a person, it must meet the minimal condition of *embodiment*, as well as several specific conditions that we explore in the following section.

> interfaced to facilitate skillful coping with environmental contingencies that are significant to the collectivity as such. (Huebner 2013: 199)

To illustrate this concept on a common example, cognitive architecture refers to a set of distinct elements or subsystems that are interconnected in a manner that enables, for instance, unhindered movement across various types of terrain and the ability to circumvent physical obstacles to reach point B from point A. Along with the elements essential for physical movement – e.g. wheels, tracks, legs, paws, etc. – the architecture also comprises a sensory apparatus and a processor that processes information obtained from the environment to adapt its behavior and overcome obstacles.

The aforementioned type of cognitive architecture pertains to the most basic forms of intelligent (i.e. adaptive) behavior. However, for an entity to be considered a person in the full sense of the term, it is clear that additional elements are necessary. Therefore, the question arises as to what constitutes an adequate cognitive architecture for an entity, whether biological or not, to be considered a person. We propose that an adequate cognitive architecture for an entity (whether biological or artificial) to be considered a person can be represented by the following set of conditions:

a) The entity must have propositionally structured intentional states, such as beliefs, desires, hopes, and intentions.
b) The entity must possess at least some form of sensory apparatus.
c) The entity must possess at least some means of interaction with the environment.

These three conditions frame our understanding of personhood. We should clarify, however, that we do not take them to be exhaustive or definitive, and that some philosophers maintain that at least some of these conditions are not even necessary. Specifically, according to Olson's animalistic view, the concept of personhood does not necessitate the presence of any of these conditions.[8] Yet, in contrast to Olson's view, we argue that condition (a), for instance, pertains to the capability of having intentional states, and that its inclusion in our proposal is self-evident, as an entity devoid of such capabilities cannot be considered a person. In other words, we think it is evident that the necessary condition for an entity to be considered a person, both in a biological and artificial context, is the ability to possess properly structured intentional states; namely, an entity cannot be considered a person if it is unable to possess beliefs, doubts, or the ability to question its own existence as a person.[9]

8 This conclusion is a direct implication of Olson's position that a person is numerically identical to her biological organism, even in instances where the individual is in a persistent vegetative state (see Olson 2000: 7–10).

9 Whether a person should possess additional mental states – e.g. sensations, affections, etc. – is a separate and complex issue that falls outside the scope of this paper.

An in-depth examination of the rationale for the implementation of conditions (b) and (c) will be presented with regards to individuals in a persistent vegetative state. For the purpose of clarity, it is important to note that the persistent vegetative state is defined as a condition of extremely severe brain damage in which the patient exhibits no observable behavior, despite appearing to be awake (Cranford & Smith 1979: 203). Now, studies conducted by neuroscientist Adrian Owen (2006) have revealed a method of successful communication with these patients, which suggests that these individuals still meet the criteria outlined in conditions (a), (b), and (c) which are necessary for an entity to be considered a person. The case of Owen's patients exemplifies the importance of each of the conditions in the cognitive architecture, and also illustrates the possibility of a "hybridization" of personhood, as their cognitive architecture includes elements that are not necessarily a part of their physical organism. Furthermore, the case of Owen's patients suggests that the determination of an adequate cognitive architecture should not be restricted to the biological organism alone.[10]

The question of whether an entity that satisfies only one or two of the conditions mentioned above can be considered a person is a complex and nuanced one. Specifically, can an entity that possesses an appropriate network of propositionally structured intentional states, but lacks both sensory apparatus for contacting the environment and means of interaction with it, be considered a person? In addressing this question, it is important to consider the practical implications of the concept of a person, such as legal, ethical, and epistemic considerations. These implications require at least some form of contact or interaction with the environment. Therefore, it can be argued that an entity without any means of contacting or interacting with the external world cannot be considered a person.[11]

However, it should be noted that there are certain borderline cases, such as the example of individuals in a persistent vegetative state, as presented by Owen (2006), where the status of a person may be dependent on the possibility

---

10 It is important to note that this example is not intended to provide a definitive answer to the question of whether individuals in a permanent vegetative state should be considered as persons. Rather, it serves as an illustration of the complexity and nuances of the concept of personhood and the importance of considering different perspectives and evidence when examining this question. Furthermore, it is important to keep in mind that this is a complex and ongoing debate in the field, and more research and studies are needed to fully understand the implications of cognitive architecture and personhood.

11 This thesis unambiguously derives from the classical understanding of the concept of person, which traces its origin to Locke (1975: 2.27.26), and according to which the term "person" is primarily a "forensic term", i.e. a term that has ethical and legal significance. Specifically, we argue that an entity which satisfies condition (a) alone but not conditions (b) and (c) may still belong to the class of *moral patients* (i.e. the class of entities towards which actions may be evaluated in moral terms), although it certainly does not belong to the class of *moral agents* (i.e. the class of entities whose actions may be evaluated in moral terms). For further information on this distinction, see McPherson (1984).

of establishing some form of rudimentary interaction with them. This suggests that the concept of a person is not a rigid one and may depend on the specific circumstances and context. Furthermore, it can be argued that the conditions proposed by our architectural view also determine the temporal boundaries of a person. Specifically, an understanding of the cognitive architecture of a person allows for the specification of the conditions that determine when a person begins and ceases to exist.

With this in mind, let us now examine how the most common examples of borderline cases of personhood would be characterized according to our proposed framework:

- *Is a fetus a person?* The fetus does not meet the requirements for personhood as outlined in conditions (a)-(c). Although it has a rudimentary sensory apparatus, it does not possess the necessary mental states or propositional intentional structures to be considered a person.[12]
- *Is a newborn a person?* A newborn may have some limited abilities to interact with the environment, but it does not possess the necessary mental states or propositional intentional structures to be considered a person.
- *Is a three-year-old child a person?* A child of this age possesses the necessary cognitive architecture to be considered a person, as outlined in conditions (a)-(c).
- *Are primates (e.g. chimpanzees) persons?* Some research suggests that chimpanzees and gorillas possess the cognitive architecture necessary for personhood as outlined in conditions (a)-(c).[13]
- *Is the chatbot "Sophia" a person?* Sophia may have advanced conversational abilities and sensory apparatus, but it does not possess the necessary mental states or propositional intentional structures to be considered a person.[14]

---

12 This point contradicts the position on fetuses put forth by Olson (1997: 96).

13 For more information on the philosophical argument that at least some animals are persons, see Aaltola (2009), and Rowlands (2019). It is important to emphasize, however, that our conclusion is a matter of philosophical debate. Specifically, some philosophers challenge the capacity of non-language-using animals to form propositional attitudes, which would render them non-compliant with condition (a), and therefore not qualified as persons. For further information on these discussions, see Davidson 1982; Dreckmann 1999; Fellows 2000.

14 It is acknowledged that some advanced robots currently possess the capability to experience tactile and auditory sensations. For the purposes of this argument, it is assumed that the chatbot "Sophia" possesses such capabilities as well. However, it is noteworthy that the communication abilities of Sophia, while highly developed, are based solely on the functioning model of a pure algorithm, which is not sufficient in and of itself to meet the requirements outlined in the condition (a) for personhood. The possibility of meeting this condition through the use of a sufficiently developed algorithm is not explored in this discussion. The reality is that at present, Sophia does not meet this

- *Is Andrew, the protagonist from Isaac Asimov's science fiction novel, The Bicentennial Man, a person?* Andrew certainly possesses the cognitive architecture necessary for personhood as outlined in conditions (a)-(c), and thus, according to our proposal, he would be considered a person.
- *Is a human being in a permanent vegetative state a person?* According to the results of Oven's experiments, these patients possess the cognitive architecture necessary for personhood as outlined in conditions (a)-(c).
- *Is a human being who is in a coma or deceased a person?* A person in a coma or deceased does not meet any of the conditions for personhood as outlined in conditions (a)-(c).[15]

As we can see, our three conditions effectively address all the borderline cases by avoiding the drawbacks of the program and hardware views while, at the same time, retaining their positive aspects. Additionally, these conditions classify these cases in a way that aligns with our common-sense intuitions. On the one hand, our architectural perspective avoids the radical anti-psychological stance of Olson's animalism and his interpretation of the hardware view, according to which psychological characteristics of a biological or artificial entity do not constitute a necessary condition for determining personhood. On the other hand, the psychological/program view is frequently criticized for its pronounced virtuality, which renders the concept of personhood *ephemeral* and fundamentally *immaterial*. By emphasizing the necessity of material or physical realization and the conditions under which an entity must possess the ability to interact with its environment, our architectural concept effectively addresses this limitation.

The rationale for adopting this position is that it allows for a separation of function and implementation, with the program remaining at the level of functional description and the hardware remaining at the level of implementation. In other words, we take it that by transitioning to the level of cognitive architecture, the main theoretical challenges for the hardware and the program view can be avoided. This is achieved by considering cognitive

---

condition in practice and, as such, does not possess the cognitive architecture necessary for it to be considered a person.

15 This point directly contradicts the thesis advanced by David Mackie (1999). In agreement with our assertion, Olson presents a well-developed argument to demonstrate that an organism does not persist after death. What is commonly referred to as a "dead body" is simply a remnant of the organism that cannot be identical to any organism that was once alive (Olson 2000: 142–53). Similarly, Leonard Sumner (1976: 153) also advocates for the perspective that death represents the cessation of existence. This understanding has numerous significant philosophical antecedents. For instance, Epicurus articulates a perspective in which death represents the end of our existence. This thesis, which is present in Epicurus' philosophy and also advocated by Olson, Sumner, and numerous others, is known in contemporary literature as the "termination thesis".

architecture as referring to physical structures, such as the presence of a sensory apparatus and the processing of internal and external representations, without requiring that these structures be neural or a part of a biological organism. In this way, our position provides a consistent definition of the concept of personhood that applies to both biological and artificial contexts.

6. Concluding remarks

In conclusion, we have explored the connection between cognitive architecture and personhood in this paper. After delving into the topic, we have come up with a set of criteria that must be met for an entity, whether it is biological or artificial, to be considered a person. These requirements include having propositionally structured intentional states, possessing some form of sensory capabilities, and having the ability to interact with the environment. We have argued that these criteria provide a framework for comprehending the concept of personhood, and the case of individuals in a persistent vegetative state, studied by Owen, serves as a prime example of the importance of these conditions and the possibility of personhood being a "hybrid". Owen's research suggests that these individuals still fulfill the outlined criteria, proving the significance of cognitive architecture in determining personhood. In the end, we are aware that further research is necessary to fully grasp all of the important implications of cognitive architecture in regards to personhood, for this complex concept demands a more profound understanding, which cannot be fully achieved within the confines of this study alone.

References

Aaltola, E. (2008). "Personhood and animals." *Environmental Ethics* 30, 175–193.

Asimov, I. (1990). *The bicentennial man and other stories*. Gollancz.

Baker, L. R. (2000). *Persons and bodies*. Cambridge University Press.

Blatti, S., & Snowdon, P. F. (2016). *Animalism: New essays on persons, animals, and identity*. Oxford University Press.

Boden, M. (1990). *The philosophy of artificial intelligence*. Oxford University Press.

Chisholm, R. (1991). "On the simplicity of the soul." *Philosophical Perspectives* 5, 167–181.

Cranford, R. E., & Smith, H. L. (1979). "Some critical distinctions between brain death and the persistent vegetative state." *Ethics in Science and Medicine* 6, 199–209.

Davidson, D. (1982). "Rational animals." *Dialectica* 36, 317–28.

Dreckmann, F. (1999). "Animal beliefs and their contents." *Erkenntnis* 51, 597–615.

Duch, W., Oentaryo, R. J., & Pasquier, M. (2008). "Cognitive architectures: Where do we go from here?" *Frontiers in Artificial Intelligence and Applications* 171.

Fellows, R. (2000). "Animal belief." *Philosophy* 75, 587–599.

Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT Press.

Huebner, B. (2013). *Macrocognition: A theory of distributed minds and collective intentionality*. Oxford University Press USA.

Langley, P., Laird, J. E., & Rogers, S. (2009). "Cognitive architectures: Research issues and challenges." *Cognitive Systems Research* 10, 141–160.

Lazović, Ž. (2009). "Neurofilozofija na delu: filozofske pouke neuroloških defekata." *Theoria* 52, 115–125.

Lepore, E., & Pylyshyn, Z. (1999). *What is cognitive science*. Wiley-Blackwell.

Maslin, K. (2001). *An introduction to the philosophy of mind*. Polity Press.

Mackie, D. (1999). "Personal identity and dead people." *Philosophical Studies* 95, 219–242.

Milojević, M. (2018). *Metafizika lica*. Institut za filozofiju, Beograd.

Noonan, H. (2019). *Personal identity*. London: Routledge.

Olson, E. (1997). "Was I ever a fetus?" *Philosophy and Phenomenological Research* 57, 95–110.

Olson, E. (2000). *The human animal: Personal identity without psychology*. New York: Oxford University Press.

Olson, E. (2019). "The metaphysics of artificial intelligence." In Guta, M. P. (ed.), *Consciousness and the ontology of properties*. Routledge, 67–84.

Owen, A. M., Coleman, M.R., Boly, M., et al. (2006). "Detecting awareness in the vegetative state. *Science* 313, 1402.

Pollock, J. (1989). *How to build a person*. MIT Press.

Putnam, H. (1964). „Robots: Machines or artificially created life?" *Journal of Philosophy* 61, 668–91.

Rowlands, M. (2019). *Can animals be persons?* Oxford University Press.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Pearson.

Searle, J. (1980). "Minds, brains and programs." *Behavioral and Brain Sciences* 3, 417–24.

Snowdon, P. F. (1990). "Persons, animals, and ourselves." In Gill, C. (ed.), *The person and the human mind: Issues in ancient and modern philosophy*. Oxford: Oxford University Press.

Sokić, M. (2020). "Lični identitet i teorija psihološkog kontinuiteta." *Theoria* 63, 87–104.

Sumner, L. W. (1976). "A matter of life and death." *Noûs* 10, 145–171.

Sutton, C. S. (2014). "The supervenience solution to the too-many-thinkers" *Philosophical Quarterly* 64, 619–639.

Turing, A. (1950). "Computing machinery and intelligence." *Mind* 59, 433–60.

Janko Nešić
Institute of Social Sciences
Belgrade, Serbia
jnesic@idn.org.rs

Original Scientific Paper
UDC 165.242.1:165.62

I AM MINE: FROM PHENOMENOLOGY OF SELF-AWARENESS TO METAPHYSICS OF SELFHOOD

Abstract: *I aim to show that, contrary to standard deflationary or eliminativist theories of the self, we can argue from the phenomenology of pre-reflective self-awareness for the thesis that subjects of experience are substances. The phenomenological datum of subjectivity points to a specific metaphysical structure of our experience, that is, towards the substance view rather than the bundle or the minimal self view. Drawing on modern philosophical accounts of pre-reflective self-awareness, mineness and (self-) acquaintance, I will argue that a subject is aware of being the one individual who has many experiences and that it is revealed to the subject that it is the bearer of experiences and their unifier. The subject is present in pre-reflective awareness and known as the subject of experiences, and even this minimal self-awareness gives us reason to favour the substance view. Thus, one can demonstrate how the debates on the phenomenology of pre-reflective self-awareness and the metaphysics of selfhood intersect.*

Keywords: *Pre-reflective self-awareness, substance, bundle, self, subject of experience.*

How adequate unto itself
Its properties shall be
Itself unto itself and none
Shall make discovery —
Adventure most unto itself
The Soul condemned to be —
Attended by a single Hound
Its own identity.
This Consciousness that is aware, Emily Dickinson

1. Introduction

When it comes to answering the question of the nature of selfhood, most contemporary philosophers of mind have gravitated towards deflationary and eliminativist accounts.[1] As a response to such strategies, I will put forward phenomenological arguments for the claim that the self is a substance. I aim to

1 Even those who are sympathetic to non-physicalist accounts of consciousness are deflationists about selves and subjectivity. See Strawson (2009), Chalmers (2015). For a *no-self* view, read Metzinger (2003).

 DOI: 10.5937/BPA2336067N

provide arguments from the phenomenology[2] of pre-reflective self-awareness for the thesis that the subject of experience is a kind of substance – that the phenomenological situation points to a specific metaphysical structure of our experience in favour of the substance view.[3] If a subject is pre-reflectively aware of being the one individual who has many experiences (it is revealed to the subject that it is the bearer of experiences, their unifier and individuator), then this goes to show that the subject is, indeed, a substance. I will discuss if all the conditions for this can be satisfied in the course of the paper. It will be argued that the debates on the phenomenology of pre-reflective self-awareness and the metaphysics of subject/self are closely related.

The overall plan of the paper is the following. In Section 2, I will present the main competing metaphysical theories of selfhood. In Section 3, I survey modern conceptions of subjectivity and mineness, showing what the best theories of pre-reflective self-awareness are and how subjectivity affords us self-acquaintance needed to defend the substantival nature of subjects. Section 4 presents arguments that we are plausibly acquainted with ourselves as subjects and that we have self-acquaintance. Section 5 discusses the "bundle" theory of the self, and Section 6 "the minimal self" theory of Zahavi. Section 7 is dedicated to the phenomenological defence of the substance view of the self, based on the findings from previous sections.

2. Theories of the self

Which theories of the self are on the metaphysical table? There are two traditional camps: the substance theory and the bundle theory of the self. I will add a third, recent position – the minimal/phenomenal self. We can define these three main categories of metaphysical theories about experiencing subjects (selves) in the following manner.

If the self is understood as a bundle (*The Bundle View*), a subject is individuated by experiences (identity conditions of persons are specified in

2 Here, I do not refer to the Phenomenological tradition and philosophical method but to the phenomenality of experience, the "what-it's-likeness" of experiences (and the subjectivity of experience), though many crucial ideas on the nature of pre-reflective self-awareness come from phenomenologists like Husserl, Merleau-Ponty and Michel Henry.

3 Thesis that experiencing subjects are bearers of experiential properties in which they are instantiated should not be a trivial claim. Nida-Rümelin (2017, 56) warns of this possible trivialization and shows that it comes from a simple argument from instantiation. Such an argument doesn't do justice to the philosopher's claims about experiencing subjects. It is not enough to infer from the truth that an experience is an instantiation of an experiential property in something, that "something" must be an experiencing subject. The truth about the subject being a bearer of experiential properties has to come from reflection on the nature of subjects and experiential properties. And we can only know something about the nature of experiences from our phenomenology. So, the fact that we are after a different solution gives more weight to the argument of this paper.

terms of relations between mental states) which are psychological modes – the subject is just experiences and their relations; selves are collections of properties. (e.g., Hume, 1978; Parfit, 1984; Dainton, 2008).[4] Selves reduce to experience.

In case the self is a kind of "minimal subject" (*The Minimal View*), it is identical with the *subjectivity* (with *a feature*) of experience. For example, Dan Zahavi (2014) defends *experiential minimalism* in which the *for-me-ness* or *first-personal givenness* dimension/aspect of phenomenal consciousness *is* the *minimal experiential self*. He is explicit on this: "More precisely, the claim is that the (minimal or core) self possesses experiential reality and that it can be identified with the ubiquitous first-personal character of the experiential phenomena" (Zahavi 2014, 18). This modern approach is a descendant of theories from the Phenomenological tradition. The "minimal/thin" theories are deflationist about the self and have some similarities with the bundle view.[5]

What does it mean to say that the subject of experience is a *substance*? It would be to claim that an experiential subject is a kind of metaphysical entity that acts as a bearer of experiential properties, upon which experiential properties are instantiated, and is not itself a property (e.g., P.F. Strawson, 1959; Chisholm, 1969; Lowe, 1996; Nida-Rümelin, 2018). It is that which endures among many changing experiences. Hence, the instantiations of experiential properties in subjects are types of events, namely *experiences*. For the subject involved, it is like something to undergo the experience because experiential properties are such that there is something it is like to have them.

If the self or subject is a substance (*The Substance View*), it metaphysically unifies and individuates experiences (experiential properties) as their bearer. Lowe writes (1996, 9): "selves or persons as *substances* — that is, as enduring

---

4 A modern variant of the bundle theorist is dynamical *self-organization*. See Metzinger (2003, 2011).

5 Galen Strawson's (2009) *sesmet* theory of self (*sesmet* being the acronym for "subject-of-experience-as-single-mental-thing") could also be called *thin self theory*, with the difference being that the experiential self is *diachronically* persistent in Zahavi's account. Guillot (2017, section 3.5.2) discusses minimalism about the self: "the self, or at least a form of selfhood (the "minimal self" or "core self"), is identical either with experience, or with some part or intrinsic property of experience", citing Zahavi and Williford (2015) as proponents, while saying that Strawson's view "bears a resemblance to this type of minimalism". The case of Strawson's *sesmet* seems to be peculiar. Strawson holds that selves are objects (2009, 298) because of their "strong unity", though he is distancing himself from the traditional notion of an individual substance. With every experience there is a subject of experience, experiencing involves a subject, and this is a metaphysical and a phenomenological claim for him. Sometimes he claims that there is an identity between the subject and experience. In a sense, Strawson could be understood as claiming that selves are substances (in a new, more relaxed sense of *substance/object*), though not enduring substances. If this is the case, his theory would belong in the first category. I do not have the space to analyze Strawson's view in more detail.

bearers of successive states and in no way reducible to mere successions of those states".

Experiencing subjects are fundamental.[6] With these metaphysical theories in mind, we can now turn to the phenomenology of subjectivity in the endeavour to properly understand what has broadly been called *the subjective character* of consciousness (and other key notions found in these debates, like those of *self-awareness* and *mineness*). Self-consciousness can be understood in many ways, but we need the fundamental type, *pre-reflective self-awareness*, the best candidate for the type of awareness that provides a grasp of the nature of the experiencing subject. When we arrive at a clearer understanding of what pre-reflective self-awareness is, then we need to show that acquaintance and self-acquaintance are possible and, thus a way to the nature of the experiencing subject.

3. Self-awareness and mineness

It has become common in contemporary analytical philosophy of mind to hold that consciousness has a *subjective* as well as a *qualitative* aspect, that there is a difference between what an experience is like and it's being like something for its subject (e.g., Levine, 2001; Kriegel, 2009).[7] This subjective dimension of consciousness has been understood in different ways by different philosophers. This side or aspect is sometimes called for-me-ness, me-ness, mineness, first-personal givenness or simply subjectivity.[8] Mineness and subjective character are sometimes meant to stand-in for the *pre-reflective self-consciousness* of the phenomenological tradition. Many phenomenologists and philosophers of mind maintain that something like self-consciousness in the pre-reflective and pre-conceptual sense exists. This

6 E. J. Lowe held "the ownership view", and has argued against neo-Humean and neo-Lockean theories of personal identity, though he did this from a proper grasp of one's self-knowledge. In *Subjects of experiences* he concluded that: "The self must be conceived of as having the status of a substance vis-a-vis its thoughts and experiences – they are 'adjectival' upon it (are 'modes' of it, in an earlier terminology), rather than it being related to them rather as a set is to its members." (Lowe 1996, 195). "Property-instances are ontologically dependent entities, depending for their existence and identity upon the individual substances which they characterize, or to which they 'belong'" (Lowe 2006, 27). It can be noted that Lowe found the doctrine of 'bare particularity', that there is a 'substratum' or 'bare particular' supporting property-instances, indefensible. He maintained that the modes belong to the individual substance itself. "I contend that modes are 'particular ways objects are', and as such are ontologically dependent upon objects in a much stronger sense than, according to a trope theorist, any trope can be ontologically dependent upon other tropes in a bundle of compresent tropes" (Lowe 2006, 97). For more on substances and bundles, see Lowe (1998, 2006).

7 From Nagel (1974) to Zahavi (2014); two *dimensions* of experience.

8 Also, *inner awareness* (Kriegel, 2009; Farell and McClelland, 2017).

feature of consciousness is also called *pre-reflective self-awareness*.[9] The self-awareness in question is not of the cognitive kind, deployed in I-thoughts, but minimal, non-reflexive.

Now, mineness could be a misleading term. Are experiences phenomenally presented as mine? One will not, perhaps, find any feature or property of the experience, a stamp or a mark, that would say that *it is mine*. Even if there is such a feature in phenomenal consciousness, does it commit one to the existence of a subject of experience? There is much imprecise talk when the matter of subjectivity is concerned. Some of the confusion behind the use of such notions was cleared by the work of Siewert (2013), Nida-Rümelin (2014, 2017) and Guillot (2017). They have shown how we should work towards developing better and more accurate concepts based on our phenomenology.

We need to be careful when using the umbrella term "subjective character" because it can designate essentially very different things. Nida-Rümelin (2014, 2017) shows there are three interpretations of "subjective character": *basic intentionality, primitive awareness and awareness of basic intentionality*. What she calls *awareness of basic intentionality* is *pre-reflective self-awareness*. She argues that awareness of basic intentionality cannot have the structure of basic intentionality and is not itself experiencing.

Philosophers of consciousness since the Phenomenological tradition have pointed out that there is something special about pre-reflective self-awareness. Such awareness is very hard to pin down. It is such that a subject is never an object in its own stream of consciousness. It is not as if one is turning "the mind's eye" inward. Nida-Rümelin (2014) would say that the awareness of basic intentionality (self-awareness) is not itself an experience that exhibits basic intentionality. The subject is not presented to itself "as an object".[10]

In a recent paper, Guillot (2017) proposes that *the subjective character* refers to several distinct notions that are being confused by certain authors:

---

9 When I say *self*-awareness, I mean *awareness of the self* and not *awareness of awareness* as in „higher-order" and "self-representational" theories of consciousness. For a good discussion on this distinction and on the relation between phenomenality and self-consciousness, see Siewert (2013). Pre-reflective self-awareness could be present universally, in every conscious creature that is a subject of experience, and in every episode of experiencing.

10 Apart from Nida-Rümelin (2014, 2017), such a "non-objectual" view of pre-reflective self-awareness in modern philosophy of mind is also to be found in Zahavi (2014). However, it is arguable what exactly Zahavi means by "non-objectifying form of self-consciousness". He is alternating between *subject-self-consciousness* and *state-self-consciousness*. Siewert contends that "presence" or "givenness" of experience, how experience is phenomenally "for me" or "mine", should be understood as a kind of self-awareness: "think about the way the viewpoint of the looker is implicit in how things look" (Siewert 2013, 31). A form of self– awareness is built into the experience: "marginal" *awareness of oneself as a looker.*

for-me-ness (a relation of awareness between a subject and an experience of hers), *me-ness* (a reflexive relation of awareness a subject has to herself) and *mineness* (a relation of awareness between a subject and a fact that it owns the experience, the fact of ownership) and all these are about relations of awareness between *a subject* and *its experiences* (Guillot 2017, 32).[11] These are distinct properties not to be conflated. Guillot uses case studies of *depersonalization syndrome* and *thought-insertion* to support her tripartite framework. There is no *prima facie* entailment between these notions. Mineness and me-ness are not universally present, though for-me-ness seems to be.[12] Guillot argues that the property of mineness can be lacking in certain cases, like schizophrenic "thought-insertion" cases. Schizophrenic patients may lack mineness, awareness of ownership, but they could still have self-awareness (*me-ness*).[13]

4. Self-acquaintance

If we have self-awareness and this awareness affords self-acquaintance, we are on our way to having a better understanding of the metaphysical nature of the self. First, it needs to be proved that *acquaintance* with one's *self* is possible. Modern acquaintance theory comes from Russell and denotes the closeness and intimacy of the subject to her experiences (experiential properties of consciousness). The consequence of this closeness is that the nature of experiences is *revealed* to the subject. The most likely candidates for the things we are directly aware of or acquainted with are *experiences*, that is, experiential properties of consciousness. Other candidates are the *subjects* of those experiences.[14] Modern proponents of the acquaintance approach to introspective or phenomenal knowledge are Gertler (2012), Goff (2015, 2017), Horgan and Kriegel (2007), and Nida-Rümelin (2007, 2016).

---

11 Howell and Thompson think that these notions are about *Phenomenal Me-ness*, for which there are two conditions: *The Phenomenal Condition* – Phenomenal me-ness must make some contribution to a subject's total phenomenal character. *The Representational Condition* – Phenomenal me-ness must in some way present or refer to the self (Howell and Thompson 2016, 4).

12 What Zahavi has in mind when he talks about *mineness* is actually the first notion, namely, *for-me-ness.* In Guillot's interpretation, *me-ness* is what Nida-Rümelin calls *pre-reflective self-awareness*.

13 There are several possible views on the prevalence of these features in consciousness. Farell and McClelland (2017, 4–5) see three options: Universalism (inner awareness, as they call it, is present in all non-reflective experiences), Typicalism (not present in atypical cases) and Absentism (never present). Three forms of inner awareness that they make distinct, following Guillot, are also: for-me-ness, me-ishness and mineness. In their terms, Nida-Rümelin 'is a universalist about for-me-ness. She also seems to be at least a typicalist, and perhaps a universalist about something similar to me-ishness and something similar to mineness' (2017, 12).

14 Which are generally left out of the acquaintance discussion.

Brie Gertler's (2012) approach is explicitly defended as more "modest" than Russell's theory, though it is a descendant of this theory. Gertler's *acquaintance approach to introspective knowledge* is expressed by the claim that we can sometimes *directly* grasp our experiences, and in such situations, we form phenomenal concepts and introspective judgments about our experiences. These make up our *knowledge by acquaintance*. As Gertler formulates it, the main thesis is that in grasping experiences, phenomenal reality "intersects" with the epistemic (Gertler 2012, pp. 94–95). Three conditions need to be met for something to be a judgement of introspective knowledge. Such introspective judgements are directly tied to their truthmakers. For their justification, they only depend on the subject's conscious states. They are more justified than empirical judgements (Gertler 2012, 100). The Acquaintance claims that the gap between *epistemic appearances* and *phenomenal reality* is sometimes filled. There is also a *metaphysical* claim here since phenomenal reflects the metaphysical reality: judgements are directly tied to their truthmakers – experiences, and these are experiential events.[15]

Acquaintance is the thesis that our intimacy with experience puts us in the position to know the real nature of the experience – the nature of the thing acquainted with is revealed to us. Goff calls it the *Real Acquaintance* and defines it as: 'A psychologically normal subject can come to know the real nature of one of her phenomenal qualities by attending to that quality' (Goff 2015, 124). A closely related thesis he proposes is: '*Phenomenal Certainty*: A psychologically normal subject is able to put herself into a situation in which, with respect to one of her phenomenal qualities, she is justified in being certain that that quality is instantiated (where to be certain that P is roughly to believe with a credence of 1 that P)' (2015, 124). That is to say, phenomenal knowledge is completely infallible. When one has an experience, there can be no doubt that one has it, that the given experiential property is instantiated (translated to the *framework of experiential properties* terminology). Goff's thesis of *Phenomenal Certainty*, which is not only implied by *Real Acquaintance*, is encountered once again as explained by *Real Acquaintance*, because it is a very plausible thesis in its own right. In Goff's theory, it is coupled with *Phenomenal Insight*.[16]

---

15 Note that in Russellian acquaintance the relation of acquaintance is between a subject and a thing (sense-datum). In modern accounts it is a relation between an introspective judgment and its truthmaker.

16 "*Phenomenal Insight*: We have rich a priori knowledge concerning our phenomenal qualities." (Goff 2015, 128). Goff defends "*Phenomenal Transparency*: Phenomenal Transparency is the thesis that phenomenal concepts reveal the essence of the states they denote. According to Revelation, when a person attends to a token conscious state under a direct phenomenal concept, the complete nature of the type to which it belongs is apparent to her; this entails Direct Phenomenal Transparency: the thesis that direct phenomenal concepts are transparent" (Goff 2017, 108).

If there is acquaintance, then the Revelation thesis is true. To know the nature or essence of a property (phenomenal property P) is to know what it is for the property to be instantiated. If I have a sensation of purpleness, then I know that an experiential property of being phenomenally presented with purple is instantiated in me.

Why is acquaintance important? Because when one is acquainted with something, the *nature of the thing is revealed*. Why wouldn't the same hold for self-awareness, not just awareness of the experience? Self-awareness is the awareness that the self has of itself that *is direct and immediate*, unmediated (Horgan and Nichols, 2016). So, it would be natural to expect that we are thus acquainted with ourselves, that we have self-acquaintance. And if this is "real self-acquaintance" then the nature of the self is revealed to us in acquaintance. If the self can satisfy these requirements, then it can be claimed that we have self-acquaintance in addition to acquaintance with experience (properties).

It seems plausible that for one to argue that the self is a substance and that the self knows this from its experience, one would need a premise that would state the possibility of a subject being *acquainted* with oneself. To know its own nature, a subject must have the proper ability to know this nature to have access to that nature. However, one need not expound on ambitious notions of acquaintance in order to do so. I can be wrong about the precise content of some experiences, but I cannot be wrong that I am having some experiences right now, whatever they might be.

One is especially acquainted with oneself because (pre-reflective) self-awareness or self-presence is so intimate it is immediate and direct (unmediated). Is *self-acquaintance* as plausible as acquaintance with experiences, or do we need additional arguments for it? Many believe that when we have direct awareness of something, then we are acquainted with it. If we have direct awareness of the self, if there is self-awareness, then the self or subject is acquainted with itself. Duncan (2015) has argued that the self passes, what he calls, *The Doubt Test*, which is a test for acquaintance with something. *The Doubt test* can be found in theories from Descartes' to Russell's, but also in modern theories, like Gertler's (2012) approach and Horgan and Kriegel's (2007). This test states that if we cannot doubt the existence of something being presented to us in awareness, then we are acquainted with it. We can doubt that the object that is producing my experiences of it exists, but I cannot imagine any sceptical scenario which would make me doubt that I have any experiences in the first place.

Duncan points out that in the case of an acquaintance with our experiences, we are in the position to be aware of their essence, but it also seems to some philosophers that this is not the case in self-awareness. There could exist an *asymmetry* between experiences and the self. Acquaintance with the self is only partial, revealing only some aspects. But there is no real

asymmetry here. The self is as much (directly) revealed as the experience. Both experiences and the self could have hidden aspects unrevealed. What is presented, though, is being directly aware of.

Though *prima facie* it may look as if there is a difference between acquaintance with experiences and acquaintance with the self in the sense there is an *appearance/reality gap*, Duncan argues that this is not the case. Experience is as it seems, and the self's properties could be misleading in the way experience would not be. But the self and the experience are on par with this; the same could be said for experience. There is no appearance/reality gap with some aspects of the self, like me being the subject of a certain experience. There is no such gap between seeming and being the subject of my experiences and the same for 'occupying a certain perspective'[17] of a subject (Duncan 2015, 2546). In both cases, we cannot doubt that there is something phenomenally in awareness (of experience and the self). Therefore, I cannot doubt that I have some experiences and that it is me that is the subject of these experiences. If this is the case, according to a plausible Doubt test, then we can be acquainted with both our experiences and ourselves.[18]

5. What is it like to be a bundle?

To vindicate my claim that we can infer from the phenomenology of pre-reflective self-awareness that subjects are substances, I would like to demonstrate that there is no phenomenological proof for the bundles of experiences view of the self. Phenomenological differences in awareness of a bundlist and a substantivalist should be highlighted.

---

17 Talk of "perspectives" can also be misleading. See Nida-Rümelin (2017, Section 10) for discussion.

18 Russell contemplated the possibility of self-acquaintance but was cautious since he considered it a difficult question, but he admitted that it is *probable* for the acquaintance of selves to occur, "though not certain" (Russell 1912, 50–51). Russell says that there is acquaintance with two things in relation (self and its sense-datum), if one is acquainted with his acquaintance with a sense-datum: "Self-acquainted-with-sense-datum". He contends that even to know the truth of being acquainted with a sense-datum, we need to be acquainted with the "I", the self. There is a striking likeness between what Russell says about self-acquaintance and Guillot's formulation of mineness, awareness of the fact of ownership (that a subject has an experience). There are those that argue that we are not directly aware of ourselves but indirectly through being aware of our experiential states (Chisholm, 1969). Chisholm would argue that to be "acquainted with the self as it is" just is to be "acquainted with the self as it manifests itself as having qualities" (1969, 21). In support of the opposite claim, take into account what Horgan and Nichols write about *the zero point*: "The self that is present in consciousness directly and without the mediation of a self-representation— the me that is experientially present via the for-me-ness of consciousness—is directly present in experience" (Horgan and Nichols 2016, 148). They use slightly different terminology, that is, instead of pre-reflective self-awareness, they use "non-representational self-presence" or just "phenomenal subjectivity".

A bundlist would deny that there is self-awareness or mineness in any of the forms defended earlier. What one needs to do is to anticipate the would-be bundlist response to substantivalist arguments: they could say that the *core bundle* plays the role of the subject (essential properties of the bundle) and that this fact is indiscernible from substantivalist phenomenology, that is, it feels the same phenomenologically as being a substance. It would appear that the bundlist has a more economic theory of the self because it posits only one category.

One such bundle theory of the self is Barry Dainton's *phenomenal self* theory.[19] By postulating something like a phenomenal *background*, a bundlist could explain self-awareness, it is claimed. Still, I would argue that, on the ground of acquaintance in awareness, the difference in metaphysics produces a difference in phenomenology.

Dainton understands mineness as a *meish* quality to experience (Dainton 2008, Ch. 8) and asks if it exists in phenomenology. As it was argued in earlier sections, such understanding of mineness and self-awareness is ill-conceived and misleading. The phenomenal background has an inner component, and this consists of "bodily experience, thoughts, memories, imaginings, and emotions", that is, of *experiences*. The inner phenomenal background creates (constitutes) the feeling of being me or you, the ambient "sense of self". The natural intimacy of "mineness" is gained when a new experience is incorporated in this background. Slors and Jongepier (2014) argue that the mineness of experience is a product of the *external* structure of experience in their *reductionist coherentist* account. It would appear that what they are considering as mineness is very different from what Zahavi has in mind or what we find in Guillot (2017). Although Slors and Jongepier agree with Zahavi that thoughts have first-personal givenness, the mineness they are defending has nothing to do with how it is usually conceived. It is the sense of familiarity of coherence of a certain experience with other background experiences. It is similar to Guillot's third property of mineness. Also, there is no real phenomenal datum to it, and Slor and Jongepier are very explicit about this, saying that there is "an absence of a further experiential feature" (Slors and Jongepier 2014, 194).

That the experience is mine is explained by the *co-consciousness* of this experience with the inner component of the phenomenal background. This background is the phenomenal *I* present in consciousness (Dainton 2008, 243). He points out that Parfit has also advocated "the reductionist view of

19 Donnchadh O'Conaill writes how Dainton "has developed a sophisticated version of the bundle theory", distinct from the classic bundle theory of Hume, one in which the subject is a bundle of *capacities for experiences*, and not a bundle of experiences themselves (O'Conaill 2019, 1–2). O'Conaill argues in his paper that Dainton's co-consciousness, as a relation of experiential togetherness, presupposes a common subject of the experiences and that the identity-conditions of experiential capacities cannot be specified without their subjects.

our *sense of self*".[20] Dainton thinks that there is no special awareness of the self as a thing, awareness of the subject as a subject, as that which is experiencing something, *the experiencer*. Apart from this being metaphysically problematic, it seems to me that it is phenomenologically unjustified. Usually, Dainton's C-system and the *phenomenal self* theory are attacked from metaphysics. I think that Dainton's theory should be criticized from a different (and arguably more plausible) understanding of mineness as self-awareness, as a real *sense of self*, that is as pre-reflective self-awareness of the subject as a subject of experiences.

6. Minimal self

What is it that we cannot doubt and that we are acquainted with when it comes to ourselves as subjects of experience? We cannot doubt, at least, that we have some experiences (experiential properties instantiated) and that it is us (me, you) who have those experiences, although we can be wrong about what exactly they are, what some of their aspects are. One could deny that in self-acquaintance we are presented with an individual essence, that in self-awareness it is disclosed to one that one is a specific, individual subject that bears the mark of uniqueness.

With his "minimal (experiential) self" theory, Zahavi tries to defend a third, middle-way position, between substance and bundle views of the self (Zahavi 2014, 18):

> The phenomenological proposal can be seen as occupying a middle position between two opposing views. According to the first view, the self is some kind of unchanging soul-substance that is distinct from an ontologically independent of the worldly objects and conscious episodes it is directed at and of which it is the subject. According to the second view, there is nothing to the consciousness apart from a manifold or bundle of changing experiences. There are experiences and perceptions, but no experiencer or perceiver. A third option is available, however, the moment one realizes that an understanding of what it means to be a self calls for an examination of the structure of experience, and vice versa.[21]

---

20 Billon uses depersonalization cases to challenge Dainton's inner background theory of mineness, because depersonalisation can affect *all* conscious states, even those in the background. This suggests to Billon that mineness is explanatory prior to co-consciousness. Billon contends that in normal cases, we have unimpaired basic self-awareness, but this self-awareness cannot inform us on the nature of the self (Billon 2017, 6).

21 The experiential self of Zahavi has temporal extension and is something that can be shared by many (changing) experiences, although there may be interruptions of the stream of consciousness (unconscious episodes of sleep and coma). This sets Zahavi's theory apart from the bundle view, though he does not posit an extra self (as a substance)

The self is seen as a feature or function of the givenness of experience, as a dimension of experience that defies both elimination (in a bundle) and inflation to a substance. One could also call this the "*thin subject"* view.

Following what has been said in the discussion of Guillot (2017), it has become clear that in order to have a phenomenological and a metaphysical claim about the subject of experience, one needs something more than just *for-me-ness* of experience. Self-awareness is thus needed for this (me-ness). So, to have any introspective knowledge about the subject, we first need to have self-awareness. There is, perhaps, no mineness as a feature of the experience, but there is "mineness" as awareness between a subject and the fact of ownership. If there was just something like *for-me-ness* in experience (which is what Zahavi usually assumes to be *mineness*), this would be insufficient to support the subject as substance claims.

Guillot criticizes Zahavi's position by showing that he moves from an epistemic to a phenomenal and a metaphysical thesis, "from the 'self-manifestation' of experience (*for-me-ness*) to a phenomenal access to the self (me-ness)" (Guillot 2017, 50). Zahavi makes an illegitimate move based on an unjustified assumption of equivalence, because he conflates *for-me-ness* with *me-ness*, and ends up claiming that a property (for-me-ness as a quality of experience) *is* the "minimal self".

In the next section, I will discuss what we can learn about the nature of the subject of experience from pre-reflective self-awareness (me-ness).

7. Individual nature of the self

When one has peered into the essence of the subject (the self) and has been acquainted with oneself, one is aware that he is a thing that has experiences. One philosopher who argues for this kind of revelation of the subject in self-awareness is Nida-Rümelin (2017 75):

> But even before such conceptualization we are aware of ourselves as 'uniting' simultaneous and subsequent experiences. And if to unite simultaneous and subsequent experiences partially characterizes our own nature as experiencing beings, then this means that we are, in pre-reflective self-awareness, aware of ourselves as belonging to that particular ontological category; we are thus aware – in pre-reflective self-awareness – of ourselves as subjects in the following substantial sense: our nature is present to us in such self-awareness in a phenomenologically manifest way.

to account for the diachronic unity and personal identity, either. Still, some kind of *awareness of diachronicity* in pre-reflective self-awareness is preserved (Zahavi 2014, 77). This is explicitly stated by Zahavi: "Whether the same experiential self is present in two temporally distinct experiences depends on whether the two experiences in question partake in the same dimension of mineness or for-me-ness" (2014, 72).

What this means is that in pre-reflective self-awareness we are aware of ourselves as entities (things) that unite experiences and are their bearers; the owners of such and such experiences. If this is our nature as subjects (or, at least, a partial aspect of our nature), then we are aware of this aspect or characterization of our nature, we are aware of ourselves as unifiers of experiences. This is the "general concept" we have of an experiencing subject and it is based on pre-reflective self-awareness.

Self-awareness based conceptualization of the fact that "simultaneous instantiations of experiential properties are instantiated by one and the same subject" (Nida-Rümelin 2017, 76) is *nature-revealing*. What this conceptualization reveals is *the simple view*.[22] This could also be put thus: to be aware of oneself as the one who stays the same in changing experiences that one has. It is the same subject who has all the simultaneous and past experiences and is engaged in actions.

According to Nida-Rümelin, in the self-awareness based understanding of synchronic unity, self-awareness pre-reflectively gives us the nature of ourselves as subjects, that we are unifiers of experiences. If we could conceive of a reverse case: that there is a causal connection between experiences (or a *co-consciousness* relation that Dainton posits) that makes them simultaneously mine, but I do not grasp it. That is, the situation is due to the causal facts, but I do not conceptualize them, then the concept of synchronic unity is *opaque* in Goff's (2011) terms and this does not seem to be right. If this was the case, then my self-awareness based understanding of my synchronic unity does not reveal to me what it is for me to have simultaneous experiences and Nida-Rümelin rightly warns that this is an unacceptable scenario. Self-awareness based understanding of synchronic unity is nature revealing and there is self-acquaintance (and self-revelation).

Exactly in pre-reflective self-awareness, if this analysis is right, we are aware of ourselves as the one who unites the experiences, this is part of our

---

22 The simple view states that simultaneous experiential properties are instantiated in one subject. Nida-Rümelin (2017, Section 14) goes on to argue that pre-reflective self-awareness also gives us an understanding of our own diachronic unity, of what it means to have experiences at different moments belonging to the same subject. With it we get the *simple view about diachronic unity* and the *simple view about transtemporal identity* of subjects (Nida-Rümelin, 2012). The simple view or *non-reductive view* with respect to personal identity and diachronic unity was also advocated by E. J. Lowe (1996). Lowe writes: "Moreover, the self's substantial simplicity is in no way incompatible with its manifest psychological complexity, though that simplicity does help to explain its psychological *unity*. The simplicity of the self is seen to imply that its diachronic identity — its persistence through time – is irreducible and ungrounded, and hence criterionless" (1996, 10). Zahavi discussed the issue of diachronic unity in his *experiential self* account, and concluded that such self has temporal extension even before obtaining narrative capacities and that "our pre-reflective self-consciousness includes some awareness of diachronicity" (Zahavi 2014, 77).

nature revealed, we are "aware of being the one single individual who has those properties at once" (Nida-Rümelin 2017, Section 13), of being the individual who has many simultaneous experiences. All these experiences are united because they belong to that one individual.

A closely related issue is that of *the phenomenal concept* of the subject/self. How could we make sense of such phenomenal concepts? How can there be any concepts of the subject in pre-reflective self-awareness? Nida-Rümelin (2017) tries to account for this with the "general concept" of the experiencing subject. Although friends of the Acquaintance/Revelation thesis gladly defend phenomenal concepts of experiences (experiential properties or phenomenal qualities), the same is not easily said of phenomenal concepts of subjects. There is very little literature on the topic today and substantial work is to be done in order to defend the plausibility of such phenomenal concepts.[23]

Let us now ask the important question: in order for this phenomenological argument to work, should a subject be aware of the fact of ownership or is pre-reflective self-awareness enough? If all traits of substantival nature are revealed in self-acquaintance, then it can be inferred, very straightforwardly, that the self is revealed to be a substance. Perhaps, the property of *mineness,* in Guillot's terms, that is an awareness between a subject and a fact that it owns the experience (where ownership is revealed), would be the most persuasive phenomenological evidence. Still, it could turn out that this property is not essential for the subject and could be absent in pathological cases.

One could argue that, given the definition of a substance, three conditions need to be met – the subject/substance is the *bearer, unifier and individuator of experiences*. Regarding the third metaphysical requirement for something being a substance, I find it hard to understand what *individuation* would "look" like in our phenomenology, if it is phenomenally present at all. If only the first two traits are revealed – that the subject is the bearer and unifier of experiences and not their individuator – then we might need a further argument.[24]

If the most plausible accounts of pre-reflective self-awareness and mineness are taken, as discussed in previous sections of the paper, arguably, some substance-like traits are revealed. Even if phenomenology does not justify the claim that the subject is aware of all the needed traits, enough data may be present in awareness to conclude that the subject is a substance.

---

23 Guillot argues for one "phenomenal model" of the concept of the self (I-concept), which is grounded in *cognitive phenomenology*, specifically in *the phenomenology of intellection* (e.g., Guillot, 2016).

24 If we have a *transparent* phenomenal concept (terminology of Goff, 2011) of the subject, it is such that the whole nature of its referent is revealed. However, if only a part of nature is revealed, we would have a *translucent* concept.

In pre-reflective self-awareness, an experiencing individual is aware of itself as an *individual*, that is, aware of its own *individual nature*[25], and this nature is very different from the one revealed in acquaintance with experience.

Recall what was said in Section 4. If one finds Goff's *Phenomenal Insight,* or a similar claim*,* plausible, and one is acquainted with one's self and with one's experiences, as it is claimed, then one knows that the subject is something essentially different from experience (self-awareness presents essentially different content from the content of awareness of experience) and one would not confuse these two. One could then use *Phenomenal Insight* to give support to the present argument. If there is acquaintance and if the essence of the self and the experiences is revealed, the subject should be able, on the basis of that acquaintance, to see the distinction between the self and the experience.

It would seem that something else would be known in self-acquaintance if the self is a bearer of properties, then what would be known if the self was a property or an aspect itself? This difference in facts can be found in different revelations of bundlists and substantivalists. This is seen in the difference between pre-reflective self-awareness (awareness of oneself) and awareness of experiential content ("objects" that are phenomenally presented to the subject). Here, we should take into consideration what was said earlier about the specific nature of pre-reflective self-awareness, something that makes it unique (being non-objectual awareness). That there is a difference in the contents of these awarenesses was stressed in Section 3. I find that this gives us an additional argument in support of the claim that the self is a substance.

---

25 One could explain the specific content of pre-reflective self-awareness with a reference to a *haecceity* ("thisness" or "individual essence") at the heart of the conscious individual. Could there be something like a haecceity of the subject of experiences? One version of the view was held by Swinburne (1995). His position is that only *conscious* beings have haecceities and can *grasp* those haecceities. 'The property of being me, if it exists, might indeed be called a 'perspectival' property-a property which something has in virtue of being thought of or grasped from a particular 'point of view' (its own)' (Lowe 2003, 88). Rosenkrantz (1993) defended the plausibility of haecceities in every object and argued that a person can grasp its own haecceity, that each individual is acquainted with himself, though haecceities of physical objects are ungraspable. Following the same intuition, Nida-Rümelin has defended that conscious individuals have a *non-descriptive* individual nature (Nida-Rümelin, 2012). One does not need to understand essences as properties. If there is a nature or essence of pain, it is not a further property that the property of pain has (Goff 2015, 126). Positing haecceities has intuitive appeal in the case of conscious individuals (subjects). Although a proponent of the *no-self* approach, Metzinger writes about a 'distinct phenomenology of singularity, a non-sensory phenomenology of 'thisness'– for example, in the phenomenology of meditation, but also in bodily self-consciousness. If we look closely enough, we can discover the phenomenology of primitive 'thisness' in our own subjective experience. It is particularly distinct in certain non-conceptual layers of self-awareness' (Metzinger 2011, 282). See Lowe (2003) for a related discussion on individuation.

8. Conclusion

I argued that the phenomenology of pre-reflective self-awareness gives weight to the metaphysical claim that subjects of experience are substances. To back this up, I argued from phenomenology that selves have experiences as instantiated experiential properties of which it is a bearer. It needs to be indicated that the *Acquaintance* needed in my argument is very minimal – what is only needed is that the subject is present in awareness and known *as the subject* of experiences, not that we have some knowledge of it as a substance. We are only aware that there *is* a subject. Therefore, I restrict my claim to saying that we are aware of our experiences and the subject of those experiences and that this gives support to the substance theory. The goal of this paper was to show that the phenomenological situation of pre-reflective self-awareness favours the substance view of selfhood, not to show what kind of a substance the self is, nor what precise theory of substances should be endorsed.

Acknowledgement I would like to thank Martine Nida-Rümelin and Jacob Naito for many helpful discussions on the topics of this paper, as well as the audience at the EXRE Colloquium (University of Fribourg). I am grateful to Donnchadh O'Conaill for useful comments and remarks on a previous draft of the manuscript.

References

Billon, A. (2017). Basic Self-Awareness. *European Journal of Philosophy,* 25 (3), 732–763.

Chalmers, D. (2015). Panpsychism and panprotopsychism. In Alter, T. and Nagasawa, Y. (eds.), *Consciousness in the Physical World: Perspectives on Russellian Monism*, pp. 246–277. Oxford: Oxford University Press.

Chisholm, R. (1969). On the observability of the self. *Philosophy and Phenomenological Research,* 30(1): 7–21.

Dainton, B. (2008). *The Phenomenal Self,* Oxford: Oxford University Press.

Duncan, M. (2015): We Are Acquainted With Ourselves. *Philosophical Studies* 172/9: 2531–2549.

Farrell, J., McClelland, T. (2017). Editorial: Consciousness and inner awareness, *Review of Philosophy and Psychology* 8: 1–22.

Gasser, G. & Stefan, M. (2012). *Personal Identity. Complex or Simple?* Cambridge: Cambridge University Press.

Gertler, B. (2012). Renewed acquaintance. In D. Smithies & D. Stoljar (Eds.), *Introspection and consciousness*, pp. 89–123. Oxford: Oxford University Press.

Goff, P. (2011). A posteriori physicalists get our phenomenal concepts wrong. *Australasian Journal of Philosophy* 89(2): 191–209.

Goff, P. (2015). Real acquaintance and physicalism. In P. Coates & S. Coleman (Eds.) *Phenomenal Qualities: Sense, Perception, and Consciousness*, pp. 121–141. Oxford University Press.

Goff, P. (2017). *Consciousness and Fundamental Reality,* New York, USA: Oup Usa.

Guillot, M. (2016). Thinking of oneself as the thinker: the concept of self and the phenomenology of intellection. *Philosophical Explorations*, 19(2): 138–160.

Guillot, M. (2017). I Me Mine: on a confusion concerning the subjective character of experience. *Review of Philosophy and Psychology: Special Issue on Consciousness and Inner Awareness,* 8 (1), 23–53.

Howell, R. & Thompson, B. (2017). Phenomenally mine: in search of the subjective character of consciousness. *Review of Philosophy and Psychology,* 8 (1), 103–127.

Horgan, T., & Kriegel, U. (2007). Phenomenal Epistemology: What Is Consciousness That We May Know It So Well?, *Philosophical Issues* 17 (1), 123–144.

Horgan, T., & Nichols, S. (2016). The zero point and I, In S. Miguens, G. Preyer and C. B. Morando (eds.), *Pre-Reflective Consciousness: Sartre and Contemporary Philosophy of Mind,* pp. 143–176. London: Routledge.

Hume, D. (1978 /1740). *A treatise of human nature*, Oxford: Oxford University Press.

Kriegel, U. (2009): *Subjective consciousness: A self-representational theory*, Oxford: Oxford University Press.

Levine, J. (2001): *Purple haze: The puzzle of consciousness*, Oxford: Oxford University Press.

Lowe, E. J. (1996): *Subjects of Experience*, Cambridge: Cambridge University Press.

Lowe, E. J. (1998): *The Possibility of Metaphysics: Substance, Identity, and Time,* Oxford: Clarendon Press.

Lowe, E. J. (2003): Individuation. In Michael J. Loux & Dean W. Zimmerman (eds.), *The Oxford Handbook of Metaphysics*, pp. 75–95. Oxford: Oxford University Press.

Lowe, E. J. (2006): *The four-category ontology: A metaphysical foundation for natural science*, Oxford: Oxford University Press.

Lowe, E. J., (2012): The Probable Simplicity of Personal Identity. in G. Gasser & M. Stefan (eds.), *Personal Identity: Complex or Simple*, pp. 137–155. Cambridge: Cambridge University Press.

Metzinger, T. (2003): *Being No One,* Cambridge: MIT Press.

Metzinger, T. (2011): The no-self alternative. in S. Gallagher (ed.), *The Oxford Handbook of the Self*, pp. 279–296. Oxford: Oxford University Press.

Nagel, T. (1974): What Is It Like to Be a Bat? *Philosophical Review*, 83 (4), 435–450.

Nida-Rümelin, M. (2007): Grasping phenomenal properties. In Torin Alter & Sven Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, pp. 307–388. Oxford: Oxford University Press.

Nida-Rümelin, M. (2012): The non-descriptive individual nature of conscious beings. In G. Gasser & M. Stefan (eds.), *Personal Identity: Complex or Simple*, pp. 157–176. Cambridge: Cambridge University Press.

Nida-Rümelin, M. (2014): Basic intentionality, primitive awareness, and awareness of oneself. in A. Reboul (ed.), *Mind, Values and Metaphysics. Philosophical Papers dedicated to Kevin Mulligan – Volume 2*, pp. 261–290. London: Springer.

Nida-Rümelin, M. (2016): The experience property framework – a misleading paradigm. *Trends in philosophy of language and mind, special issue of Synthese*, 195 (8): 3361–3387.

Nida-Rümelin, M. (2017): Self-Awareness. *Review of Philosophy and Psychology: Special Issue on Consciousness and Inner Awareness* 8 (1): 55–82.

O'Conaill, D. (2019): The identity of experiences and the identity of the subject. *Philosophical Studies*, https://doi.org/10.1007/s11098–018–01226–4.

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

Russell, B. (1912): *The Problems of Philosophy*, Oxford: Home University Library.

Siewert, C. (2013): Phenomenality and self-awareness. in U. Kriegel (ed.), *Phenomenal Intentionality*, pp. 235–258. Oxford: Oxford University Press.

Slors, M. V. P., Jongepier, F. (2014): Mineness without Minimal Selves. *Journal of Consciousness Studies,* 21 (7–8), 193–219.

Strawson, G. (2009): *Selves: An Essay in Revisionary Metaphysics*, Oxford: Oxford University Press.

Swinburne, R. (1995): Thisness. *Australasian Journal of Philosophy*, 73 (3): 389–400.

Strawson, P. F. (1959): *Individuals: An Essay in Descriptive Metaphysics*, London: Methuen.

Williford, K. (2015): Representationalisms, Subjective Character, and Self-Acquaintance. In T. Metzinger & J. M. Windt (Eds), Open MIND: 39(T). Frankfurt am Main: MIND Group.

Zahavi, D. (1999): *Self-awareness and alterity: A phenomenological investigation*, Evanston, Illinois: Northwestern University Press.

Zahavi, D. (2014): *Self and Other: Exploring Subjectivity, Empathy, and Shame*, Oxford: Oxford University Press.

Miljana Milojević
Faculty of Philosophy
at the University of Belgrade
miljana.milojevic@gmail.com

Original Scientific Paper
UDC 165.12(091)
159.923.2:165.12

THE NOTION OF A PERSON*

Abstract: *The aim of this article is to clarify the content of the concept "person" as it figures in philosophical debates about personhood and personal identity. In order to do so, I will look at both specific philosophical problems that ask for a clear definition of this notion, as well as at the history of this concept's formation, and try to motivate the specific assumptions that are tightly connected to it.*

Keywords: *Personhood, Personal Identity, Substances, Properties*

1. Self, personality, human, or what persons are not

The concept *person* is an everyday concept that seldom needs clarification in common usage. Mike, John, Jamey and Nina are persons. They are human beings that surround us with their specific personalities, thoughts, feelings, rational choices and legal rights. If we look at the definitions of different positive laws, we will encounter similar determination – natural persons are human beings coming to existence by birth and ceasing to exist with death. Also, the common usage of the term "person" will usually coincide with the usage of the term "self". A quick search on the internet will give us definitions of the "self" as "person's essential being" or similar, which makes "self" in common usage just a different way to refer to an individuality of a single person. On the other hand, a common usage of "personal identity" will point us to different kinds of personalities and/or idiosyncrasies of a human's psychological profile.

Nevertheless, in the philosophical literature these terms are carefully kept separate and are seldomly used for mutual definitions. In other words, meanings and extensions of "human", "personality", "self" and "person" in philosophy do not necessarily coincide. For instance, if we look at the relation of *self* and *person* which are tightly connected in common use, we will find mostly words of warnings about their connection in the philosophical literature. As Heersmink (2020) notices, there are weaker and stronger

* This research was financially supported by the Ministry of science, technological development and innovation of the Republic of Serbia as part of the funding of scientific research at the University of Belgrade – Faculty of Philosophy (contract number 451-03-47/2023-01/ 200163).

concepts of self than those of a person. For instance, minimal self is identified with the sense of ownership of one's experiences and actions, and does not suffice for constitution of personhood, while narrative self "is constituted by the content of [person's] self-narrative, and the traits, actions, and experiences included in it are, by virtue of that inclusion, hers" (Schechtman 1996: 94; cf. Heersmink 2020: 3), and thus surpasses the content of a *person*, at least the philosophical one, as we shall see in the remainder of the paper. In that sense, we should avoid identifying persons with selves.

Also, philosophers will often warn us not to conflate the philosophical notion of "personal identity" which refers to the problem of tracking persons through time, with psychological notion of a personal character or personality. Namely, in everyday communication the meaning of "personal identity" will almost exclusively be explained in terms of different character traits, inclinations, talents, set of desires, and similar. A similar connection of persons and personalities is made in most written laws with respect to personal rights. Among the rights that a person has over their body, law protects under the term "personal rights" a series of "personality aspects" such as person's talents, honor, and reputation. "Personal identity" in philosophy, on the other hand, almost always refers to the problem of diachronic identity, or the problem of determining what needs to remain the same in the changing entity which is a person to be called the same person over some given period of time. Such criteria almost never refer to personality traits and instead focus on metaphysical theories of persistence and sortal identity.

As we shall see, philosophical notion of a *person* is not even equated with that of a *human*. This is because philosophical theories of personhood look to determine the essential properties of a person, and *being human* is usually seen only as an accidental property of a person. Namely, we can imagine persons which are not humans, like intelligent Martians, or artificial intelligence that is equal or surpasses that of a human. Thus, it is frequently claimed that persons are only contingently human because the essential properties of a person, whatever they are (for now we will have to settle with a placeholder), are had only by humans in our known environment. On the other hand, the notion of personhood is seldom debated in everyday life, and we can live happily most of our lives devoid of skepsis with just equating persons with humans. Every now and then a conversation about someone's pet or a smartphone assistant will come about, and the question of their potential personhood will be often lively examined. Those debates will settle on one or the other side with arguments being previously thrown in both directions. Some of the reasoning against the idea of non-human persons can be based on "they do not have a soul", "they do not have feelings (in case of AI or both)", "they are disembodied (in case of AI)", etc. claims, while the opposite can be defended based on "they do have feelings (in case of animals or both)", "they are equally intelligent (in case of strong AI)", "they have a sense of self", etc. claims. What is clear is that our everyday notions are not equipped for settling questions about personhood. Nevertheless, we are

coming to our main question now and that is whether is philosophy better off in this manner? Just a brief look at debates about personhood and personal identity informs us about a great disagreement between different theories about what persons in fact are. So far, we got to the point of claiming that persons are not selves, personalities nor humans, but what are they then? Also, even if we cannot settle on specific essential properties of persons, can we at least settle on the unique philosophical concept of a person?

Contrary to everyday usage of a term "person", in several philosophical debates "person" is used as a theoretical term and based on a particular theory of personhood and personal identity "person" will have different meaning and different extension. According to these theories persons will usually be only contingently, and not essentially, human and they will be identified with various sets of properties – sometimes biological, sometimes physical, and sometimes psychological. Also, some of these theories will be reductive and try to identify persons with some other type of known entities, and some of them will be non-reductive – constructing persons as new kinds of special entities.

Given this diversity in individuating and identifying persons, someone might say that trying to say something generally accepted about the philosophical concept of a person must be a futile endeavor (see for instance, Travis 2015; Naffin 2011). Nevertheless, all the different theoretical concepts of a person have a common core distinct from the one of its commonsense counterparts, one which is defined by a set of specific questions that have to be resolved, at least this is what is going to be argued. The subject of this article is precisely this shared content of different philosophical *persons*, that are sometimes identified with humans, suitable biological organisms, souls, psychological entities, etc. Thus, the task before us is to determine which problems and which questions led us in a search for identifying conditions and identity criteria of persons, or in other words, which common interests of philosophers led them to construct different theories of personhood and personal identity which will in turn reveal the meaning of a philosophical concept of a person.

2. How did we come to construct the notion of a person?

We can identify at least three important historical sources of our contemporary notion of a person[1]. Namely, in the ancient concept of *prosopon* (anc. gr. *πρόσωπον*), Roman law's *persona*, and theological debates about the God's nature, we can find three important elements of the concept of personhood – functional, legal, and dialogical.

1 A large part of Section 2., one that is dedicated to historical reconstruction of the concept of a person, is based on Chapter 1 of *Metaphysics of Persons* (Milojevic 2018:17–27). Some sentences might be translations of the text from this Chapter, especially those referring to historical facts.

The Greek concept *prosopon* can probably be identified as the historical origin of the modern concept of a *person*. It is a direct source of the Roman concept of *persona* and holds some of the essential aspects that we associate with persons today. Nevertheless, the starting formation of the concept was quite innocuous, and it did not have anything to do with philosophy, law nor morals. On the contrary, it was made to refer to one specific body part and afterwards it was readily appropriated in dramatic arts. Namely, "prosopon" was coined from two other words – "pros" and "ops" meaning "towards" and "eyes". Thus, the literal meaning of "prosopon" can be conceived as "a face" as it is the side of the head towards eyes, or as that which is in front of the eyes, signifying a connection to others and resonating the dialogical nature of persons (according to Vovolis 2009, p. 31). It is interesting to note, that in some languages, for instance in Serbian, the same word "lice" even today designates both a face and a person.

However, Greek "prosopon", with the development of ancient drama, changed its primary meaning into a theatrical mask. This happened with the introduction of actors into the performance of theatrical plays. Namely, at first a dramatic performer was delivering a text of the drama as himself. In other words, performers were just narrators and not actors. They were not playing a role, they were not assuming a character, instead they were just conveying the text as it was created by the author of the play. By Aristotle's testimony, it was Thespis who was the first actor that played a given role in the 6th century BC. Aristotle also writes that with Aeschylus the number of characters in a play rose to two, and with Sophocles to three. He also attributes the first use of a theatrical mask to Sophocles (*Poetics* 1449a). So, with the use of a theatrical mask, which was now called "prosopon", one actor could assume different roles at different times. In that sense, the idea of a *prosopon* can be best captured as a role or a function that a human can take – as the mask or personality that a human being assumes in a relevant context (Vovolis 2009). Thus, it is safe to say that in its origins the concept of a person did not refer to the human as a biological creature, but rather to her specific function. It is possible for one human to be or to play more than one *prosopons*, as well that she is not or does not play any *prosopons* or characters at all.

This idea of assuming a certain role was inherited by the Roman legal theory and developed in a new direction. Although Rome was established as early as in the 8th century BC, origins of Roman law are traced back to the Laws of the Twelve Tables dating to around 450 BC. The importance of looking at the conception of the legal notion of a *persona* is multifaceted. On the one side, Roman law is the basis of the contemporary European law, and it provided a framework for civil law. On the other hand, the vast majority of legal terminology comes directly from the Roman law, and the philosophical concept of a person is tightly connected to legal considerations. Namely, one of the most quoted passages about personhood nowadays is Locke's statement that

> „person" is foremost a *forensic term* "appropriating actions and their merit; and so belongs only to intelligent agents capable of a law, and happiness, and misery" (Locke 1694: II, xxvii, 26) and that it applies to "a thinking intelligent Being, that has reason and reflection, and can consider it self as it self, the same thinking thing in different times and places" (ibid.: II, xxvii, 9).

Thus, it is more than informative to take a look at how this term was introduced in legal/forensic context, and how it was subsequently used.

Roman law was being developed in at least two phases. The first one lasted until the 6th century and the codification of Iustinian, and the second phase, lasted from the 6th century (and especially from the 11th century onward with its Western Europe rediscovery) to the 18th and 19th century when different national laws started to develop independently and when Roman norms stopped to be followed (Mousourakis 2012: 1–2). The diversity and breadth of Roman law is admirable, but we are especially interested in one of its distinct parts, namely the one that deals directly with persons – *ius quod ad personas pertinet*. The importance of this part of the law is reflected in its position in an overall corpus of Roman law. If we look at the 6th century codification under Iustinian we will find the *Corpus Iuris Civilis*, and in this body of work we will find *Institutes* one of the three main works, which contained explanations for students of the codified law. *Ius quod ad personas pertinet* was described in the first of four chapters of *Institutes*. The law that was pertinent to persons determined their legal position in an overall social structure. It defined their rights, abilities and obligations, and it contained both status and family law. In that sense, it regulated both norms that were pertinent to individual persons as individuals and members of a society, but also those norms that applied to them as members of a family.

Interestingly, if we look at the text of *Institutes* we will find a determination of a person that significantly diverges from the contemporary legal notion of a natural person. Namely, today persons are considered as entities with legal capability, or as the subjects of the law – bearers of both legal rights and obligations. And this determination is not contingent, it is considered as an essential property of a person. In that sense, according to the contemporary legal theory a person cannot be just an object of a law. On the other hand, Roman law allowed that slaves are persons too. *Ius quod ad personas pertinent* did not regulate only the status and family relations of free citizens, but also it codified norms that applied to people who lost their freedom. We can conclude that with Roman law *persona* became a legal entity, a human being in its legal relations playing its legal role.[2]

2 Roman law recognized all humans as persons even if some persons were treated as pure objects of law – namely, slaves (as it was defined in Corpus Iuris Civilis, one of the three major parts of Institutes – great 6th-century codification of Roman law, performed under the orders of Iustinian I). It is interesting to notice that the treatment of slaves as non-

Lastly, Christian biblical exegesis brought to the fore one more aspect of personhood recognizable in the contemporary debate about persons. In an attempt to explain an apparent paradox of the claim that God is one and three at the same time, religious thinkers came up with a number of possible answers, and the most prominent one was that God is one substance, but three persons – *Una substantia, tres personae* – first formulated by Tertullian (*Adv. Prax.*; see also Tuggy 2016).[3] By using the method of prosopographic exegesis – interpretation of events through specific narratives of different dramatic roles or persons – Christian scholars noticed that trinity emerges in a dialogical relation of God to himself. God appears as three in the form of the Spirit who speaks, the Father to whom he speaks, and the Son of whom he speaks. (Ratzinger 1990: 442–443, see also Tertullian Adv. Prax. II 7–10) This dialogical nature of God which constitutes trinity was more widely conceived by St. Augustine as a relational nature of the deity and it was described by his famous analogies by which we can imperfectly gain knowledge about God (see Augustine 1991/ca. 400–420; Drecoll 2014). Thus, Christian scholarship deeply embedded in European culture further developed the concept of a person as a self-reflecting relation of an entity to itself – trait that is needed for attributing responsibilities to persons. If an entity does not have a discursive relation to itself there is no point in punishing it for its misdeeds or holding it accountable.

All three sources of the concept of a person are interrelated, but emphasize different aspects of persons: ancient Greek notion insists on a functional nature of personhood – a person is a role played in a certain context; Roman on its biological and legal aspect – a person is a human being in its legal relations to others; and Christian on its discursive dialogical manifestation – a person is constituted by a narrative relation to oneself and to other persons. We can track these ideas through history all the way to the ordinary contemporary notion of a person, which keeps these different aspects of a person by conceiving her as a human being who is capable of taking part in law by virtue of having a capacity of self-reflection or judgment of its own thoughts and actions.

We can conclude this section with the remark that contemporary notion of a person is what Locke conveniently termed a "forensic" notion. Persons

persons, and arguments that they are less than human, arose with the Enlightenment idea that persons must be only subjects of law, due to their legal capacity as animals with rational souls. Together with the idea that souls and animals are separate entities, e.g. that there could be an animal in a human form lacking rational soul, and the fact that some human animals were treated as only objects of law, this idea opened a way to arguments that slaves lack a number of cognitive capacities and that they are inferior to rational humans.

3 The importance of the question "'What is God?' (i.e., the God whom we encounter in Scripture); and, 'Who is Christ?'" (Ratzinger 1990: 439) was already recognized by Ignatio of Antioch (c. 35 – c. 107), but it was not before the 4th century that this question became one of the focal points of theological Christian discussions.

are those entities which can play a part in legal and moral order. The notion of a natural person from the Enlightenment period is a further elaboration of this notion according to which some entities can play this part due to their natural properties, arguably by being capable of discursive self-reflection and rationality. On the other hand, the notion of an artificial persons came as a matter of convenience where some entities were granted legal status not because of their natural properties, but because it was beneficial to grant them such status by fiat. Namely, *persona ficta* is a notion that originates in the 13th century and by decree of pope Innocent the IV. At that time monasteries were granted a legal status, or they were pronounced persons that can be financially and otherwise responsible, because the maintenance of the monasteries was put in jeopardy by monks' vows of poverty. Thus, although the positive law allows for legal fictions, those fictions are based on realities such are natural persons which are the subject of our and general philosophical investigation. The reality of persons in turn depends on the validity of law and moral, so inasmuch law and ethics are real and have a valid purpose, persons are real too.

But given that the concept of a person is so tightly connected to legal considerations why don't we follow the law and say that persons are simply people? If we remember the reasons for not adopting such an answer, they were based on hypotheticals about possible extraterrestrial creatures being deserving of a status of a person or similar farfetched possibilities. So, the question is should we be led by such hypotheticals and prolong our search for "proper" persons, or proper essential properties of persons, because persons are "only contingently people" in our immediate surroundings? Why shouldn't we instead rely on practical considerations and just proclaim that for all relevant purposes persons are human beings? Why not settle with the claim that persons are at least contingently human and possibly necessarily human? Well, besides the obvious philosophical need to clarify our conceptual landscape and to give a better foundation to both legal and moral theories, there is also an additional real practical need to revise our legal concept of a person. Namely, with the rise and the development of new technologies there are more and more pressing issues about the real nature of persons. Mentioned hypotheticals about beings which are not human but are deserving of the status of being a person are not anymore just mere hypotheticals. There are at least two kinds of entities which are not purely biological or human which deserve attention of both law and philosophy in this context, and those are hybrid entities and potential strong AI systems. Development of various technological aids and prosthetics blurred the boundaries of originally biological persons, and a number of authors argue that in some cases we can talk about extended or hybrid persons – persons that are partly constituted by highly integrated artifactual aids to cognitive and perceptual processing of these extended systems (Clowes 2020; Hongladarom 2016; Milojevic 2020; Piredda & Candiotto 2019). These blurred boundaries already started to have an effect on legal practice when it comes to deciding if a certain damage to an

artifact is just another property damage or a personal injury, like in the case of the damage of Neil Harbisson's "eyeborg" (see Milojevic 2017), but also in cases of privacy violations under the assumption that mental states can be stored on external devices (Palermos 2022).

Thus, we can conclude that investigating the question of what or who are persons has both theoretical and practical merit, and that is of utmost importance to define the core assumptions of different approaches to personhood and personal identity.

3. Questions that (mis)guide us in understanding what persons are

In order to determine what persons are several questions are usually separated. First, there is a question "What makes an entity to be a person?". Then, there is a question "What makes that entity or a person to be the person that she is?". Third, we have to ask given that the entity in question might change during a period of time "What makes this person in t_2 to be the same person in t_1?". And finally, "What are the physical boundaries of this particular person?". Thus, there are at least four questions and four kinds of criteria that need to be set, those of: 1) identification, 2) individuation, 3) persistence, and 4) embodiment. They are all equally important from legal and moral point of view as well. We need to know to which entities law and morals apply, how to differentiate these entities, how to track them through time in order to, for instance, attribute them with responsibilities for past deeds, and what physical parts are parts of a person in order to know when a person has been, for instance, injured.

All these questions can have different answers and various theories will advocate different sets of criteria for some or all of them (some theories will, for instance, consider only the criteria for persistence). Persons can be seen as reducible to different known kinds of physical, biological or psychological entities, or on the other hand as a new kind of entities. They can be individuated by various sets of properties that vary in kind. Their persistence can be determined by different kinds of continuity, for instance, bodily or psychological. Also, it can be argued that person's boundaries are boundaries of parts of organisms like brains, organisms, extended hybrid entities, disembodied souls, or else. We can notice that these questions and criteria are connected and some of them are more tightly mutually connected than the others. For instance, someone might say that if we claim that persons are humans as an answer to identification question, then we already have an answer for the embodiment question. The connection between the identification question and embodiment one is, thus, such that it seems that they are just two sides of the same coin. Nevertheless, if we identify persons with a functional kind that allows for multiple realizability, then the question

of embodiment becomes the question of realization which has separate answers. Also, individuation and persistence criteria seem to be dependent of each other, but it should be noticed that the first is concerned with synchronic identity and the second with diachronic identity conditions. In the end, it seems that whatever we chose as the nature of persons and how we identify them will dictate how we answer the other questions too. However, this is proven as weak heuristic. There are theories like Baker's (2000) which endorse psychological continuity as determining the personal identity through time, but do not claim that persons are a kind of psychological entities, rather they are in a physical sense constituted by their bodies. Thus, although connected, these questions should be kept separate, and in the literature we can find that authors most often focus either on identification or persistence.

The debate about personhood and personal identity usually follows, but is not limited to, arguments and counterarguments from two broadly construed camps – one that advocates physical and biological criteria, and the other that focuses on psychological properties of persons. It is interesting to notice, though, that such separation and conflict between these two camps could not even exist prior to Descartes's separation of body and mind. Namely, in the medieval times persons were identified with ensouled bodies or with rational souls which were substantially united with suitable bodies. Boethius and Thomas Aquinas called them, „individual substances with rational nature" or "naturae rationabilis individua substantia" (Aquinas, *Summa Theologiae* t, q. 29, a. 1, obj. 1.; also Boethius, *Con. Eut. et Nest*., ch. 3). Following Aristotelian metaphysics, a human or a person was seen as a living organism which had a soul as both her living and thinking principle. According to such a view, soul was a formative principle which gave functional organization to an organism which in that way gained the capacity for rational thinking and telling right from wrong, thus becoming a person. Also, it is extremely important to notice that according to Aristotle's hylomorphism souls as substantial forms could not exist in separation from their bodies. Because of such unity of biological and psychological in human substances, the dispute about the true nature of persons could not even get off the ground – there was only one contender for persons which united both kinds of properties in one essence. Debates about personhood during that period, thus, stayed away from the topics about persons' nature, and were mostly concerned with questions about temporal existence of persons – whether persons come into existence at conception, forty days after, or at birth (for an overview see Jones 2004). On the other hand, when the body and soul became separated in Cartesian philosophy the question of which one of these is a person became a valid one. Even today, when substance dualism is mostly an abandoned position with respect to the mind-body problem, property dualism and autonomy of special sciences perpetuate the duality of psychological and biological entities.

Let us look briefly at different prominent theories of personhood and personal identity in order to get a better grip on what persons are considered to be. We are going to consider a) brute physical fact theories, b) psychological theories, and c) constitution views of persons.

a) Brute physical fact accounts of personal identity adhere to the view that such identity is to be spelled out in terms of brute physical facts or relations and without the need to refer to psychological properties. Entities which are suitable for entering such a relation are, for instance, organisms and bodies. The identity of a body or an organism X_2 from t_2 with a body or an organism X_1 from t_1 consists in a brute physical fact or their physical continuity. If this relation holds for us too than we can say that we continue to exist if and only if our bodies or our organisms continue to exist. Our persistence is a matter of brute physical fact.

Proponents of brute physical fact view can be separated in two groups: radical ones which would claim that persons *are* bodies or organisms, and moderate ones which claim that *our* identity holds in brute physical facts, but we are not necessarily persons. In the first group we can, arguably, find Thomson (1997), Williams (1973, according to Parfit 2012), and Mackie (1999), and in the second group we find animalists such as Olson (1997) and Snowdon (1990). This separation is strictly provisional, and it certainly does not help that some authors use the term "person" ambiguously. Mackie, for example, does this explicitly and claims that he accepts Thomson's claim that there are dead persons (as the corpse is physically continuous with a living body) if we read "person" with a small "p" which does not ask for psychological endowment (Mackie 1999). Nevertheless, both groups of authors seem to be driven by insights such as the one that it seems that we exist even when we do not have certain psychological capacities or properties, when we are embryos or vegetative patients. Olson also argues that if we were to accept a psychological criterion for our persistence this would create a too-many-thinkers problem. More precisely, if what is necessary and sufficient for our persistence is a sort of psychological continuity, then we are not organisms, as we would persist in a different organism if our brain were transplanted. If this is so, and we also concede that organisms can think, then there are two spatiotemporally coincident thinkers – my organism and my psychological being which is not my organism, and I cannot tell which one is me doing the thinking.

In spite of these advantages, brute physical views face a number of problems especially the radical versions. If we claim that persons are bodies or organisms, we will get a neat ontology of persons based on the reduction to a known kind of entities, but the question would arise are they the right sort of entities (Baker 2000: 124). Baker heavily criticizes such a view because it does not offer a unique sort of criteria for tracking and identifying persons and by doing so it does not make them different from other kinds of physical entities. Also, such a view does not make a reference to person's psychological

capabilities which make her suitable for entering appropriate legal and moral relations. Additionally, the argument that is often given to defend animalism – view that we are animals, but not necessarily persons – can be turned against radical versions of the brute physical fact view. According to this argument, it is unintuitive to consider fetuses or unresponsive patients in vegetative state to be persons because they lack relevant psychological features needed for legal capacity (see Boyle 1979; Sherwin 1981; Olson 1997).

In the end, there are famous brain transplant and mind upload scenarios that put this position to the test. We can imagine my brain being transplanted in a different body, and given that I am my body according to such a view, I would remain in a donor body and a different person would get my brain. Such an interpretation of the brain transplant scenario seems implausible, and we have strong intuitions that I would wake up in a new body after the transplant. Also, given that I am my body I cannot ever have a different body. This would stop mind upload scenarios, and all other scenarios where we would be differently realized, in their tracks. Thus, according to the brute physical fact view I cannot ever be uploaded to a computer or to a robotic body, not as a matter of contingency, but as a matter of conceptual necessity.

These unintuitive consequences show us that even if these accounts can potentially individuate persons and successfully track them through time, such persons are very different from what we usually assume under the concept of a *person*. Maybe they are what Mackie (1999) termed persons with a small "p". Namely, if "person" is a forensic term as Locke (1694) claimed, and we attribute them with responsibilities, rights and other legally and morally relevant attributes, then the consequences of the defended view seem wrong and inadequate. For instance, in the case of brain swaps a person who would donate a brain of a serial killer would be still held accountable for her actions after the swap and the person who received it would be deemed innocent. So, if we doubt that persons are bodies or organisms, let us see if they are psychological entities instead.

b) Psychological view of personal identity is most famously defended by Derek Parfit (1984). According to Parfit to establish personal identity, the following has to obtain:

> (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)

Some explanations are in order. First, we should briefly establish what is psychological continuity, and then why such continuity should not take a branching form.

Psychological continuity depends on psychological connections which are causal relations between mental states. If subject S_1's mental states are caused by S_2's mental states, or if S_2's mental states are caused by S_1's mental states, there are psychological connections between S_1 and S_2. If there are multiple chains of such connections, then there is a psychological continuity between S_1 and S_2. (Parfit 1984: 206)

Introducing psychological continuity instead of physical persistence solves the problem of brain transplants – a person which is psychologically continuous with the one who committed a crime would be held responsible, meaning the one with the transplanted brain, and not the donor. Also, psychological continuity deals better with teleportation cases. Namely, if we imagine a device like the one from Star Trek and someone would have stepped into it, and she would be transported to a different place, we have a strong intuition that that person would be teleported and that she would continue to exist. Nevertheless, if a bodily criterion is a criterion of personal identity, then the transported person would be a different one than the one who entered the teleporter and whose body was destroyed in one place just to be recreated in another. The transported person is psychologically continuous with the original one, but physically discontinuous, thus psychological criterion of personal identity safeguards our intuitions about teleportation. However, teleportation and other means for differently realizing or instatiating mental states that form psychological continuity (e.g., mind uploads) open a possibility of "branching". Branching occurs when there is more than one entity that is psychologically continuous with some entity from a previous time. It can occur in different cases of fission and duplication – in cases of one-by-one brain hemisphere transplants in two different bodies, in cases of faulty teleporters which do not destroy original bodies or create two or more copies in different places, or in cases of mind uploads where the original mind retains its biological form. Branching is possible because psychological continuity is not a transitive relation, but this also makes psychological continuity different from identity relation which is transitive. In branching cases, multiple entities from a later time are psychologically continuous with an entity from a prior time, but they are not mutually psychologically continuous. So according to Parfit, in order to secure personal identity, we need to make sure that branching did not occur. The possibility of branching, thus, creates a serious problem for psychological continuity as a contender for a criterion of personal identity. Namely, the overall criteria that Parfit offers, which ask us to make sure that branching did not occur make the identity relation contingent. This, in turn, breaks the rule which Wiggins called "only a and b":

> "In notionally pursuing object a in order to ascertain its coincidence or non-coincidence with b, or in retracing the past history of b to ascertain its identity link with a, I ought not need to concern myself with things that are other than a or other than b" (Wiggins 2001: 96).

Parfit was not particularly concerned with such consequences. His most acclaimed view is, after all, that identity is not what matters in survival and other practical matters, but psychological continuity instead. It is, also, clear that relying on such a criterion of personal identity we are not coming closer to an answer to identification and individuation questions about persons. Afterall, there is "no entity, without identity" (Quine 1969). Quine's famous credo was formulated because of his dissatisfaction with identity criteria for various abstracta, and now we can be equally dissatisfied with the identity criteria for persons. Again, Parfit did not concern himself too much with the contingency of personal identity, nor with the ontology of persons and he just elliptically claimed that they are probably closer to nations or clubs (1984) and that their identity conditions are more like those given for audio-systems (1995) than for members of natural kinds. But can we after all constitute an ontology of persons whose diachronic identity relies on psychological continuity?

Thomson (1997) notices that psychological theories are usually not motivated by ontological claims and are unable to construct viable ontologies, which was one of the main reasons why she turned to the bodily criterion of personal identity. Nevertheless, there are multiple suggestions how to build psychological ontology of persons. One of them is to identify persons with brains or relevant parts of brains (McMahan 2002; Campbell and McMahan 2010). Nevertheless, such suggestion cannot meet our intuitions about teleportation, although it handles well the brain transplant cases. Also, it would betray the psychological criterion of personal identity unless we adopt identity theory of the mental – theory which is mostly abandoned and replaced with some version of functionalism about the mental. Namely, we would not be able to talk about the same person through time unless she has the same brain even if there is psychological continuity between the past and future person according to functionalism – like, for instance, in mind upload or teleportation cases.

So far, we have seen that brute physical fact views have neater ontologies but ones that do not fit our intuitions about persons nor our starting concept of a person, on the other hand psychological criterion preserves our intuitions about persons and keeps the concept of a person applicable in legal and moral contexts, but psychological views do not tell us what persons are as entities.

In the end, we are going to consider a group of theories that try to reconcile psychological criterion of personal identity with ontologies similar to those offered by some proponents of the brute physical fact view.

c) Advocates of constitution views accept a psychological criterion of personal identity and claim that although not identical, persons are constituted by their bodies. According to such a view, persons are entities spatially

coincident with organisms or bodies that have some extra properties. These extra properties make them non-identical to organisms/bodies. The most prominent advocates of constitution view are Shoemaker (1999) and Baker (2000), and their accounts differ with respect to properties that separate persons from their bodies.

Shoemaker endorses a view that properties are defined by causal roles they impart to their bearers, and by differentiating thin and thick properties he defends a view that mental properties can be properly attributed only to persons and not to animals. We can predicate "it has a cerebrum in physical state P" to both an animal and a person, but the animal and the person will have different thin properties that belong to this predicate and which present individual disjuncts of an appropriate thick physical property. It is only the person that has the appropriate mental property thanks to the appropriate causal roles that the state of the cerebrum plays in the cognitive dynamics of the psychological life of the individual that realizes the thin physical property. By defending this position, Shoemaker motivates the claim that persons are not identical with bodies, organisms nor animals, but that they are still constituted by them. Also, he answers the already mentioned too-many-thinkers argument against psychological views by excluding animals from the extension of "thinking beings".

Although, this position has many advantages – it has a plausible criterion of personal identity, the psychological one, and it gives a plausible account of embodiment of persons – it faces a number of objections. Its ontology seems artificial and *ad hoc*, produced specially to attribute relevant properties only to persons and thusly defining their essence. Árnadóttir (2010) complains that such a view implies that animals that constitute us cannot have relevant thoughts, which seems contrary to our intuitions. Furthermore, and more importantly, such a view leads to unnecessarily high standards of mentality.

Baker (2000) does not separate thin and thick properties like Shoemaker, and she does not limit mentality to persons. Instead, she introduces relevant relational properties as the properties which differentiate persons from their bodies. The example that she amply uses to illustrate the point that there could be spatially coincident non-identical objects is the one of Michelangelo's statue of David and the piece of marble in the shape of David. It was probably Aristotle who first noticed that the statue differs in some respects from the lump of clay from which it is made, and thus formulated the puzzle of non-identity of some spatially coincident objects. Baker's answer to this puzzle is that the statue has a number of properties that the piece of marble does not. The statue has properties that connect it to the art world essentially. If there was no art world the statue would not exist, or so it is claimed, even if the piece of marble would. Thus, Baker

claims that the relation between the statue and the marble is the one of constitution and not identity, because the statue has relational properties which are not properties of the marble at the same time. The same relation holds between a person and her body in Baker's view. The only question now is what are the properties that separate a person from her body? Baker argues that those are the relational properties of persons to themselves or the ability to take on the first-person perspective. In turn, Baker defines this ability in a dispositional way – object x has an ability to take on the first-person perspective at t iff it has suitable structural properties, and it has manifested this ability in a previous time, or it is in an environment suitable for developing this ability (Baker 2000: 92). This in turn enables the attribution of personhood to both fetuses and mentally disabled that protects their rights and legal and moral treatment, without the need to attribute them full-blown mental states at the same time.

Although Baker's view does not face the problem of "person chauvinism" according to which only persons can have mental states, like Shoemaker's, it is still troubled with multiple problems. For instance, given that Baker's view calls for extrinsic essential properties it is subject to the anthropic objection (see Sosa 1987, Sider 2001). This objection claims that if we allow for extrinsic essential properties there is no end in arbitrary selection of existing entities. There is no definitive answer what would make some such properties suitable for individuation and some unsuitable. In case of the statue of David would David on the table be a new entity – table statue – when it is in table circumstances? (see Wasserman 2009) Thus, if we do not answer the arbitrariness question our ontology would become overcrowded and useless.

This brings us to the end of this section. We have overviewed some of the prominent views on personhood and personal identity omitting a large number of alternative positions, some being Lewis's (1976), and Sider's (1996) perdurantist or four-dimensional accounts of persons, Nozick's (1981) "closest continuer" theory, etc. But even with this limited overview we can identify a noticeable trend. Namely, that the theories which are closer to our big "P" person concept, in Mackie's terms, one which connects persons to legal and moral issues and focus on the psychological abilities that enable attribution of relevant legal and moral attributes to them, give plausible personal identity criteria but are unable to give a consistent and unproblematic ontology of persons. On the other hand, views which have unproblematic ontologies and clear persistence criteria do not fit with our intuitions about persons and can be seen as views that give criteria of personhood and personal identity of persons with a small "p" – like the Thomson's view that allows for dead persons and persistence of de-brained ones. These insights lead us to our last section.

Concluding remarks: are persons entities or we were misguided

At the beginning of section 3. we have separated several questions that need to be answered in order to determine what persons are and how they persist. Two leading questions were the one about identification or criteria of personhood and the one about persistence or criteria of personal identity. We saw that various theories either give plausible answers to one or the other of these questions. Also, if they give unquestionable ontologies, they do not usually fit our most common concept of a person, one that identifies persons as bearers of legal and moral attributes. This leads us to the hypothesis that persons might not be concrete individuals or primary substances in Aristotle's terms in the end, and that the first and second question misled us into searching for such entities. So, if persons are not concrete individuals, what can they be? They can be properties or modes just like Locke claimed (1694), and why he claimed that a demonstrative science of morals is possible, they can be closer to artifacts as Parfit (1984) suggested, or they can be phase sortals as Olson (1997) argued. We are going to briefly look at Olson's reasoning for his claim.

In the previous section we mentioned that animalists adopt some sort of brute physical fact view with respect to our identity, but that they do not simultaneously hold that we are essentially persons. Thus, Olson frames his position as non-essentialist. He claims that we are animals, and that in some points in our lives we are persons too, but not always and not necessarily. We would like to make a similar point with a change in emphasis. Instead of focusing on our animal nature and the claim about personal non-essentialism, we would like to focus on non-substantialism of persons without any claim, for now, about what we are. Olson in "Movers and Thinkers" (chapter 2 section iii. of *The Human Animal* [1997]: 31–37) argues that the term "person" is a functional rather than a substance term, and that it stands for a phase and not a substance sortal. In order to substantiate his claim, he likens persons to artifacts, more specifically to locomotors. To make a comparison he first asks us to answer the question "What is a locomotor?" Different things such as humans, crabs, and cars are locomotors, and what makes them locomotors is their capacity to move themselves. But Olson says that we cannot answer a second valid question "What kind of thing is a locomotor? Is it a human, or a boat, or something else?" There is no structural intrinsic nature that makes locomotors to be locomotors. Furthermore, according to Olson locomotors come into existence by gaining a capacity of self-moving and come out of existence by losing this capacity. By analogy, there is nothing in the structure of a thing that makes it a person. It is rather its capacity for rationality, self-consciousness, and similar psychological abilities. Also, these capacities are had by things like humans only temporarily, after a suitable development

and before irreparable damage. Thus, Olson concludes that *being a person* is rather a phase and not a substance sortal which answers to the question "What it does?" and not "What is it?". In other words, it is closer to phase sortals like *being a child* than to substance sortals like *being an animal.*

There are several arguments against Olson's argument which is very modestly portrayed here. For example, Hershenov (2005) contests Olson's conclusion that we must be animals instead of persons because this is the most plausible substance sortal that applies to us. He does that by showing that *being an animal* is a functional term similarly to *being a person*. On the other hand, Nichols (2010) argues that many substance sortals are functional (Wiggins 2001, also allows for substance sortals to refer to functional properties). Even concepts of fundamental particles make reference to what they are capable of doing like having a certain spin. Nevertheless, even if these arguments are sound they do not show that persons are substances, they just show that reasoning from functionalism to non-essentialism is flawed, and that animalism is not the only contender even if *being a person* is a phase sortal. Olson's reasoning seems to be on the right track, but it needs amendments which would show that functions which are part of the content of the concept *person* are such that they do not constitute a substantial kind (Milojevic, unpublished).

This way we are coming full circle. We started from what persons are not in a philosophical literature, then we proceeded to survey historical origins of this notion. By identifying three sources of this notion, we singled out several aspects which are inextricably connected to the concept *person*: its functional, its reflexive, and its forensic nature. Consequently, we looked at how is philosophical research in this domain carried out, and we have identified several questions which led this research. Now we can say that some of these questions were misleading. By asking for criteria of personhood and individuation we were led to search for a kind of primary substances or concrete individuals. This search led Thomson (1997) to claim that psychological theories are inadequate and to postulate an implausible bodily criterion for personal identity. It also led advocates of constitution views to postulate questionable ontologies. While Parfit's account has all the "right" consequences that lead to unsubstantiality of persons, he never focused on ontological claims and he has not developed an account of what persons are. In the end, Olson's view clearly advocates non-substantiality of persons, but lacks a further justification of its claims. We can conclude that the vast philosophical research about what personhood and personal identity are, led us back to the historical origins of the notion of a person that emphasized that persons are roles that a human or another entity can assume, and that this concept refers to different psychological capacities that make this entity capable of law and morals. Thus, persons are best seen as functional properties and not as substances, and the further research should be directed at specifying what kind of properties they are, or at least this is what was argued for in this paper.

References

Aristotle (1996). *Poetics*. Translated with an introduction and notes by M. Heath. London: Penguin.

Aquinas, Thomas (1947). *Summa Theologiae*. Cincinnati: Benziger Bros.

Árnadóttir, Steinvör Thöll (2010). Functionalism and thinking animals. *Philosophical Studies* 147 (3):347–354.

Augustine (1991). *The Trinity*. Trans. E. Hill. New York: New City Press.

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

Boethius (1918). *Liber De Persona et Duabus Naturis Contra Eutychen Et Nestorium*. H.F. Stewart, ed., *Theological Tractates and the Consolation of Philosophy*. William Heinemann Ltd.; Harvard University Press.

Boyle, Joseph M. (1979). That the Fetus Should be Considered a Legal Person. *American Journal of Jurisprudence* 24 (1):59–71.

Campbell, Tim & McMahan, Jeff (2010). Animalism and the varieties of conjoined twinning. *Theoretical Medicine and Bioethics* 31 (4):285–301.

Clowes, Robert William (2020). The Internet extended person: Exoself or Doppelganger? *Límite: Interdisciplinary Journal of Philosophy & Psychology* 15:1–23.

Drecoll, Volker Henning (2014). Aporetic Account of Persona and the Limits of Relatio: A Reconsideration of Substance Ontology and Immutability. In Michael Welker (ed.), *The Depth of the Human Person: A Multidisciplinary Approach*. Cambridge: William B. Eerdmans Publishing Company, 186–205.

Heersmink, Richard (2020). Varieties of the extended self. *Consciousness and Cognition* 85:103001.

Hershenov, David B. (2020). Protecting Persons from Animal Bites: the Case for the Ontological Significance of Persons. *Philosophia* 48 (4):1437–1446.

Hongladarom, Soraj (2016). *The Online Self: Externalism, Friendship and Games*. Cham: Springer Verlag.

Jones, David A. (2004). *The Soul of the Embryo: An Enquiry into the Status of the Human Embryo in the Christian Tradition*. New York, London: Continuum.

Lewis, David K. (1976). Survival and Identity. In Amelie Oksenberg Rorty (ed.), *The Identities of Persons*. University of California Press, 17–40.

Locke, John (1694/1975). *An Essay Concerning Human Understanding*. In P. H. Nidditch (Ed.), *The Clarendon Edition of the Works of John Locke*. Oxford: Oxford University Press.

Mackie, David (1999). Personal Identity and Dead People. *Philosophical Studies* 95 (3):219–242.

McMahan, Jeff (2002). *The Ethics of Killing: Problems at the Margins of Life*. OUP USA.

Milojević, Miljana (2017). Extended Personhood – Rethinking Property/ Person Distinction. *Theoria* 60(4):55–76.

Milojević, Miljana (2018). *Metaphysics of Persons [Metafizika lica]*. Beograd: Institut za filozofiju.

Milojevic, Miljana (2020). Extended Mind, Functionalism and Personal Identity. *Synthese* 197 (5):2143–2170.

Milojevic, Miljana (unpublished). Persons, Substances, and Functional Properties.

Mousourakis, George (2012). *Fundamentals of roman private law*. Berlin, Heidelberg: Springer Verlag.

Naffine, Ngaire (2011). Women and the cast of legal persons. In *Gender, sexualities and law*, (ed.) Jackie Jones, Anna Grear, Rachel Fenton, and Kim Stevenson. Oxon: Routledge, 15–25.

Nichols, Peter (2010). Substance concepts and personal identity. *Philosophical Studies* 150 (2):255–270.

Nozick, Robert (1981). *Philosophical Explanations*. Cambridge: Harvard University Press.

Olson, Eric Todd (1997). *The Human Animal: Personal Identity Without Psychology*. Oxford: Oxford University Press.

Palermos, Spyridon Orestis (2022). Data, Metadata, Mental Data? Privacy and the Extended Mind. *American Journal of Bioethics Neuroscience*. doi: 10.1080/21507740.2022.2148772.

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

Parfit, Derek (1995). The unimportance of identity. In H. Harris (ed.), *Identity*. Oxford University Press. 13–45.

Parfit, Derek (2012). We Are Not Human Beings. *Philosophy 87*(1): 5–28.

Piredda, Giulia & Candiotto, Laura (2019). A Pragma-Enactivist Approach to the Affectively Extended Self. *Humana Mente* 12 (36).

Quine, Willard Van Orman (1969). *Ontological Relativity and Other Essays*, New York, Columbia Un. Press.

Ratzinger, Joseph (1990). Retrieving the Tradition concerning the Notion of Person in Theology, trans. Michael Waldstein. *Communio* 17 (3):439–54.

Schechtman, Marya (ed.) (1996). *The Constitution of Selves*. Cornell University Press.

Sherwin, Susan (1981). The concept of a person in the context of abortion. *Bioethics Quarterly* 3 (1):21–34.

Shoemaker, Sydney (1999). Self, body, and coincidence. *Proceedings of the Aristotelian Society* 73 (73):287–306.

Sider, Theodore (1996). All the World's a Stage. *Australasian Journal of Philosophy* 74 (3):433–453.

Sider, Theodore (2001). *Four Dimensionalism: An Ontology of Persistence and Time*. Oxford: Oxford University Press.

Snowdon, Paul F. (1990). Persons, animals, and ourselves. In Christopher Gill (ed.), *The Person and the Human Mind: Issues in Ancient and Modern Philosophy*. Oxford: Oxford University Press.

Sosa, Ernest (1987). Subjects among other things. *Philosophical Perspectives* 1:155–187.

Tertullian (1920 [n.d., ca. 218]). *Against Praxeas* trans. A. Souter, New York: The Macmillan Company.

Thomson, Judith J. (1997). People and their bodies. In Theodore Sider, John Hawthorne & Dean W. Zimmerman (eds.), *Contemporary Debates in Metaphysics*. Blackwell.

Travis, Mitchell (2015). We're All Infected: Legal Personhood, Bare Life and The Walking Dead. *International Journal for the Semiotics of Law – Revue Internationale de Sémiotique Juridique* 28 (4):787–800.

Tuggy, Dale (2016). Tertullian the Unitarian. *European Journal for Philosophy of Religion* 8 (3):179.

Vovolis, Thanos (2009). *Prosopon, the acoustical mask in Greek Tragedy and in Contemporary Theatre*. Stockholm: Dramatiska Institutet.

Wasserman, Ryan (2017). Material Constitution. *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2017/entries/material-constitution/

Wiggins, David (2001). *Sameness and Substance Renewed*. Cambridge: Cambridge University Press.

Williams, Bernard (1973). *Problems of the Self*. Cambridge: Cambridge University Press.

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

1

FILOZOFSKI godišnjak = Belgrade philosophical annual / editor Voin Milevski. – God. 1, br. 1 (1988)– . – Belgrade : Institute of Philosophy, Faculty of Philosophy, 1988– (Belgrade : Službeni glasnik). - 24 cm

Polugodišnje. – Glavni stvarni naslov od br. 28 (2015) Belgrade philosophical annual. – Tekst na engl. jeziku.

ISSN 0353-3891 = Filozofski godišnjak

COBISS.SR-ID 15073792