The Illusion of the Enduring Self

Katalin Balog

This paper is primarily about metaphysics; specifically, about a Cartesian view of the self, according to which it is a simple, enduring, non-material entity.[1] I take a critical look at Nida-Rümelin's novel conceptual arguments for this view and argue that they do not give us decisive reasons to uphold the Cartesian view. But in Nida-Rümelin's view, what is at stake in these arguments is not merely theoretical: the truth – and our beliefs about it – has practical consequences as well. In her view, if the Cartesian simple view of the self were false, we would have no reason to love and care for family and friends, and our special interest in our own future would be pointless. Her arguments are powerful and ingenious but ultimately, I do not find them convincing. In the last section of this paper, I will say something about the sense in which it is right and the sense in which it is wrong to think that the metaphysics of the self has broader relevance for our lives.

1. Theories of the Self

I start by a quick sketch of the two metaphysical theories of the self that will figure in this paper: the Cartesian view of the self, which – after Parfit – I will call the Simple View, and Reductionism, the alternative I prefer. They both agree about the main characteristics of the self.[2] Selves are synchronically and diachronically unified subjects of experience.[3] They are capable of self-awareness. They make decisions, are doers of deeds, and the locus of the value and dignity of a human being. Selves are evidently real: where there is experience, there is an experiencer, a subject, a self. But what is the self? The core disagreement is over whether the self is a simple, enduring, non-physical entity, or a complex phenomenon whose persistence involves connections between experience.[4]

Descartes was on the side of the simple self. He gave a philosophical formulation and defense of the prevailing Christian picture of human beings. He thought the self was an indivisible, simple, non-material mental substance whose essence is thought (by which he meant any conscious mental state, including perception, sensation, emotion, etc.), and which is a possible candidate for personal immortality – a doctrine Descartes seemed eager to uphold.[5] Thomas Reid (1785/2011) provides a classic formulation of this view:

---

[1] This view has been held by Butler (1736/1896), Reid (1785/2011), and more recently Chisholm (1976), Swinburne (2007), O'Connor (2000) and Nida-Rümelin (2010, 2011, 2017a, 2017b), among others.

[2] I will sidestep the issue of whether the self is better thought of as the whole conscious human being or as a distinctively mental entity. The latter is a viable and common conception of the self, and I will limit my discussion to it. For a detailed presentation of this conception, as opposed to the conception of the person as the whole human being, see G. Strawson (2009). But see "Persons," in P. F. Strawson (1959) for a contrary view claiming that our conception of persons, of necessity, is a primitive compound of something with both bodily and mental aspects.

[3] I use "experience" to include not only perceptual experience, but feelings, moods, thoughts, etc. Experiences are mental states such that there is something it is like to undergo them.

[4] Whether the subject even persists is a matter of controversy (Strawson 2009 argues for no persistence). I will put this issue aside.

[5] There are other conceptions of the self as a simple, enduring entity, e.g., in the Advaita Vedanta school of Hindu philosophy, the essence of the self is pure awareness, and this self is thought of as a witness to one's mental stream. I am

> My personal identity, therefore, implies the continued existence of that indivisible thing which I call *myself*. Whatever this self may be, it is something which thinks, and deliberates, and resolves, and acts, and suffers. I am not thought, I am not action, I am not feeling; I am something that thinks, and acts, and suffers. My thoughts, and actions, and feelings, change every moment: they have no continued, but a successive, existence; but that *self,* or *I*, to which they belong, is permanent, and has the same relation to all the succeeding thoughts, actions, and feelings which I call mine.

The Cartesian Simple View seems to sit well with our ordinary notion of the self. [6] It had been the prevailing view in philosophy at least until Locke[7] and Hume[8] offered their early versions of Reductionism, which holds that personal identity over time consists in memory and other psychological continuity. One of the criticisms that contemporaries levelled against Locke's memory account is that it makes immortality too cheap: God could resurrect many copies of a person carrying the same memories, making personal immortality meaningless.[9]

Reductionism denies that selves are, as Parfit (1984, 210) puts it, "separately existing entities", i.e., that they involve facts beyond psychological facts that are in principle describable without appealing to the concept "self" or "person". This is not to deny that selves exist. It is to deny, though, that they are fundamental entities that exist in addition to various mental processes existing (Parfit 1984, 252):

> Descartes' view may be compared with Newton's belief in Absolute Space and Time. Newton believed that any physical event had its particular position solely in virtue of its relation to these two independent realities, Space and Time. We now believe that a physical event has its particular spatio-temporal position in virtue of its various relations to the other physical events that occur. On the Cartesian View, a particular mental event occurs within a particular life solely in virtue of its ascription to a particular Ego. We can deny that the topography of 'Mental Space' is given by the existence of such persisting Egos. We can claim that a particular mental event occurs within some life in virtue of its relations to the many other mental and physical events which, by being interrelated, constitute this life.

Some people read Hume as denying the existence of the self altogether. But I think it is better to think of him as a Reductionist, not an eliminativist. He did not argue that there are no selves, no minds; his point was that their nature doesn't consists in being a simple, separately existing entity. He famously questioned[10] the motivation behind the Simple View:

---

putting aside these other conceptions as I will focus on Nida-Rümelin's account which fits into the Western tradition of Descartes (1641/1985), Butler (1736/1896) and Reid (1785/2011).

[6] See Bering (2006) for a fascinating account of the folk psychology of souls.

[7] *An Essay Concerning Human Understanding,* (1689/1996, Book II, Chapter XXVII).

[8] *A Treatise of Human Nature,* (1739/2000, Book I, Part IV, Section VI).

[9] Butler "Of Personal Identity", *The Analogy of Religion*, (1736/1896).

[10] Ibid.

> There are some philosophers, who imagine we are every moment intimately conscious of what we call our self; that we feel its existence and its continuance in existence, and are certain, beyond the evidence of a demonstration, both of its perfect identity and simplicity.

Instead, his own introspection reveals that

> particular perceptions…may exist separately, and have no need of any thing to support their existence…

and that we are

> nothing but a bundle or collection of different perceptions which succeed each other with an inconceivable rapidity, and are in a perpetual flux and movement… They are the successive perceptions only, that constitute the mind; nor have we the most distant notion of the place where these scenes are represented, or of the materials of which it is composed…

This is, by the way, the Buddhist view as well. I prefer to understand Buddhists as Reductionists, too, even though they call their view the "no-self" view. "No-self," here as well as in Hume, refers to the view that there is no simple, enduring self; it is not a denial that selves exist. Here, as elsewhere in metaphysics, there is a thin line between elimination and reduction. The Reductionist denies that anything answers to a certain conception of the self – the self as a simple, enduring, non-physical entity – while suggesting that there is a perfectly good sense, albeit a sense that may be farther from the commonsense notion, in which the self exists. In this understanding, the self exists as a complex of variously connected experience, "a bundle of perceptions," as Hume put it.[11]

2.Nida-Rümelin's Simple View

Nida-Rümelin provides a detailed and patiently argued account of the Simple View in a series of papers. Her central argument is that the concept of the self as a simple, enduring mental entity – *pace* Hume – is embedded in the conceptual structure of introspective self-awareness, as well as of our thinking about imaginary cases involving transtemporal self-attribution. We cannot help but use this concept when we think about experience with clarity. She then argues that if this is so, it is reasonable to accept this concept of the self as correct, there by going from conceptual observations straight to a metaphysical conclusion. Although the move from the conceptual to the metaphysical is problematic as well, here I will focus on the first part of the argument, concerning the concept of the self.

Nida-Rümelin is equivocal about the exact nature of the connection between the concept of the simple self and introspective self-awareness: is there something in the nature of introspective self-awareness that demands this concept to come on board to adequately capture experience or is it simply that this is a concept that, for other reasons, is natural for us to use in our social cognition? While her arguments seem to require the first, I will argue that at most it is the second that is the

[11] While the Simple View is anti-physicalist, a Reductionist (about the self) need not be a physicalist (i.e., a reductionist about mental states). A property dualist (e.g., Chalmers 1996), i.e., someone who holds that mental states are irreducibly non-physical, but denies the existence of non-physical, simple selves, could hold a Reductionist view about the self. Parfit himself is non-committal about the metaphyiscs of mental states. The Simple View, in other words, represents a version of dualism that is different from, and stronger than mere property dualism.

case. She misdiagnoses how our self-concept works. The issue that is my focus is a problematic move on her part from the synchronic case to the diachronic case of the unity of the self. By conflating these two cases, Nida-Rümelin mistakenly arrives at the view that when we reflect on our experience, we cannot help but understand the self as a simple, enduring mental entity.

Conscious self-awareness is essential to the self and is the basis of our understanding of the self. But as I will argue, introspection, though it reveals the nature of experience,[12] does not by itself reveal the nature of the self, much less reveal it as simple, as Hume I think rightly argues. I will maintain that, while the concept of the self as simple might come natural to people across cultures, it is a conceptual overlay, and not something that is embedded in the conceptual structure of experiential thought. In the following two sections, I will discuss Nida-Rümelin's arguments for the simple self based on the synchronic, and diachronic unity of consciousness, respectively.

2.1. The Synchronic Unity of Consciousness

Before we turn to the arguments, I would like to point out some controversial assumptions in Nida-Rümelin's treatment of self-awareness. She (2017b, 65) makes the Cartesian observation that

> experiential properties are such that having them necessarily involves being aware of having them… It is not easy to say what exactly that "awareness" consists in. … it will be helpful to have a term to refer to that special kind of awareness. I will use the term "primitive awareness" … for that purpose.[13]

She thinks (2017b, 66) we can know this to be the case by simply reflecting on what it is to have experiences:

> reflection on what it is for a property to be experiential reveals that every experiential property is such that – in a specific sense of awareness – to have that property is to be aware of having it.

This is a controversial view; and its proponents have trouble explaining what exactly this "primitive", non-reflective, albeit self-reflexive awareness really is. It is also left unexplained how reflection on experience reveals this thesis to be true. One problem with the view is that it seems to imply the idea that animals are pre-reflectively aware of their own experiences which is quite implausible; arguably, a certain level of cognitive sophistication is required for any self-awareness. Alternatively, it implies that animals do not have experiences; but that is even less plausible. The idea that we are constantly aware of our experience even when we do not consciously introspect, might be an example of the so-called 'refrigerator light illusion' (Dennett 1991), the error a child makes in thinking that the refrigerator light is always on because it's always on whenever she checks if it is.

[12] In terms of what it is like to have it. Physicalists hold that there is another sense in which introspection does not reveal the nature of experience, namely, in the sense of its fundamental composition.

[13] Sartre (1943/1956) characterizes this as "pre-reflective" self-awareness, to emphasize that this awareness does not involve any explicit further thought about experience. For versions of this view see Brentano (1874/1973) Vol.1, Book 2, Gertler (2012), Kriegel (2009), Thompson (2011), and Zahavi (2006). For higher order thought views see Lycan (2004) and Rosenthal (2002).

Although I do not think this view of pre-reflective self-awareness is essential to her arguments – since they could be reformulated on the basis of a more plausible account of self-awareness – I still want to register my disagreement here. There is a much more plausible idea, proposed by Siewert (2013), which is simply that all experience is readily available for introspection and reflection, and that such reflective states are also self-reflexive in that the experiences they refer to partly constitute them.[14] One can, for example, in reflecting on one's sensation of warmth, readily form a self-reflexive concept of that warmth, a concept that is partly constituted by that very feeling. This explains the intuition that there is a particularly intimate connection between our experience and our awareness of experience, to the point where the awareness and its object coincides. It has the added benefit that it does not require a mysterious, ever present, non-reflective self-awareness. Nor does it presuppose anything about the ontology and metaphysics of experience.

Whether ever-present or occasioned only by reflection, however, experiential self-awareness is the bedrock of our sense of self, and the foundation for any conception of self. Descartes' Cogito argument is based on the idea that each of us can secure reference to ourselves infallibly by attaching the 'I'-concept to a thought about experience[15] formed in self-awareness, as in "I think…," "I feel…," "I hear…," etc. It is in such I-thoughts that our concept of the self gets off the ground. It is unclear, however, what exactly 'I' refers to. Is it distinct from the experiences of which we can become aware of at any moment? One of the routes to the intuition that the self is distinct from experience has to do with the unity of consciousness. Here is Descartes (1641/1985 vol. 2, 19) on the unity of consciousness:

> The fact that it is I who am doubting and understanding and willing is so evident that I see no way of making it any clearer. But it is also the case that the "I" who imagines is the same "I"….Lastly, it is also the same "I" who has sensory perceptions or is aware of bodily things as it were through the senses.

Descartes famously ends up presenting arguments for the thesis that the "I" is a simple, enduring, indivisible substance. Kant agrees with Descartes about the unity of consciousness, and thinks it is the main source of the intuition of the simple self (Kant 1781-87/1998, A108):

> Thus the original and necessary consciousness of the identity of oneself is at the same time a consciousness of an equally necessary unity of the synthesis of all appearances…

Kant, however, does not think that the unity of the self across the range of conscious states implies anything about its ultimate nature. For all the intuitive appeal of the idea of a simple, indivisible substance doing the uniting, it might be a metaphysical illusion engendered by the special features of the indexical 'I' – a concept that can be applied to present, past, or future experience, without any real idea what the 'I' refers to.

Nida-Rümelin disagrees with Kant here as she did with Hume. She thinks, *pace* Hume that we can find the self in introspection, and, *pace* Kant, that we can know its nature just on the testimony of introspection. She maintains that self-awareness of present experience ratifies a conception of the

[14] For a related account of phenomenal concepts, see Balog (2012).

[15] Again, I am using 'experience' in a broad sense to include not only perception, sensation, emotion, etc. but conscious thought as well. I take the Cogito argument to work equally for 'I think', 'I feel', 'I hear', or 'I perceive'.

self as an entity that underlies – and is ontologically distinct from – experience. To argue for this, she first points out (Nida-Rümelin, 2017b, 63) that

> each of us is permanently aware of him- or herself in a pre-reflective and pre-conceptual way in any moment of his or her conscious life."

Further (Nida-Rümelin, 2017b, 75), this self-awareness reveals that experiencing subjects

> … 'unite' simultaneous experiences …How experiencing subjects are capable of doing so is something we are pre-reflectively aware of in experience. On that basis we are capable to conceptualize what it is for a subject to 'unite' simultaneous and subsequent experiences.

Her suggestion is that our very understanding of the unity of the self reveals it as simple. Putting aside once again the issue of pre-reflective self-awareness, since this does not seem essential to the argument, one can agree that introspective self-awareness reveals a unified self. But how does the a mere grasp of this unity imply that the self is ontologically distinct from experience? Nida-Rümelin offers the argument that, because there is no *conceptual reduction* of what it means for distinct experiences to belong to the same self in terms of causal-functional or any other objective properties or relations, the unity of consciousness can *only* be understood in terms of a simple experiencing subject to whom its various experiences are "presented", and which is distinct from these experiences. Her view is that it can only be a simple substance that unifies the experiences of the subject as though the self is a pin cushion and experiences pins. This is not the only view that explains unity and it does not seem to be a great explanation. For one thing, how does the self do the uniting? Merely sticking experiences onto a self does not by itself unify them. They must be related to one another in various ways, namely, they must be co-conscious. But once they are related to each other in this way, there is no need for the posit of a mental substance to do the uniting.

This is Parfit's view (1984, Chapter 12). He says the unity of consciousness "does not need a deep explanation" of any sort, causal or otherwise. "It is simply a fact that several experiences can be co-conscious or be the objects of a single state of awareness." Of course, he does not object to the idea that there might very well be a causal-functional explanation of that unity. In other words, that we are directly aware of the unity of our experience, without needing any causal-functional descriptions of it, does not mean that there could not be a perfectly good account of co-consciousness in terms of causal-functional relations. It is true that being aware of the unity of our minds does not involve any causal-functional *criteria*. But this is simply an artifact of how we can form direct phenomenal concepts of our experience by simply attending to them. Not every concept – in fact, few of them – reveal much about the nature of their referent. For all the directness of our awareness of the unity of our mind, it might very well be that what actually constitute that unity are causal-functional relations.

It is to challenge the explanation of the unity in terms of simple selves that Parfit (1984, Chapter 12) brings up divided brains. He discusses cases of brain bisection and extrapolates from the actual cases to hypothetical scenarios where people with divided brains have two equally capable, independent streams of consciousness. He argues that if this really was the case (and it doesn't seem impossible), and if such divisions could be made and unmade at will (which also does not seem impossible), Nida-Rümelin would be forced to take the awkward position that in such cases simple, immaterial subjects of experience pop into existence at each instance of division and go extinct each time the

brain hemispheres are united, without these subjects actually being persons in any ordinary sense of the term. This makes the Simple View unattractive, and the alternative more plausible as a result.

Nida-Rümelin (2017b, 58) is not convinced:

> The conclusion here drawn is conceptual: when we consider experiences …. the idea that they might occur without there being a subject involved simply does not make sense because presentation (in the sense at issue) requires there to be someone to whom something is presented.

Her response seems to indicate that Parfit's view is incomprehensible. There are two ideas here condensed into one. The first is the charge that in Parfit's view experiences might occur without a subject, which is absurd by itself; the second is that the existence of a simple subject is "required" by the fact that something is "presented" to someone. As for the first, Parfit's view would indeed be incomprehensible if he really denied that there are subjects. However, Parfit is not denying that there are experiencing subjects, or persons, or selves. He is denying that the existence of these entities involves facts *over and above* facts about experience. The accusation that her opponents are committed to the view that there are no subjects is unwarranted. As for the second idea, that experiences need some entity, ontologically distinct from themselves, to whom they are 'presented', this appears to be a misleading invocation of the spectacle/spectator dichotomy of Plato's Cave. Such images cannot be decisive when it comes to ontology. Parfit agrees that things appear this or that way to subjects but he does not think that this requires that there should be entities distinct from experiences to whom things appear. The idea that processes require a substance distinct from them, in which they can occur, has gone out of fashion in metaphysics; the object/process distinction is no longer generally considered fundamental. Experiences are no different. In the absence of further argument, Reductionism remains a viable option.

2.2 The Diachronic Unity of the Self

The argument for the Simple View that Nida-Rümelin presents in the greatest detail is the argument from the diachronic unity of the self. She holds that the I-concept is the basis of our concept of a subject or self. The most striking feature of this concept – both in its synchronic and diachronic uses – is that we apply it simply, directly, and independently of any empirical (behavioral, physical, functional, or psychological) criteria. Her crucial claim is that, despite the directness of 'I', in applying it we can form a clear and positive conception of both the synchronic and diachronic unity of the self.

My central critique is that Nida-Rümelin goes from the correct claim that, in applying the I-concept to present experience, we can form a 'clear and positive' conception of synchronic unity in that we can clearly conceptualize what it means for an experience to belong to our own self (as opposed to a qualitatively identical, but different self), to the incorrect claim that, in applying the I-concept to future experience, we can form a 'clear and positive' conception of diachronic unity, that is, we can clearly conceptualize what it means for an experience in the future or past to belong to our own self (as opposed to, say, a qualitatively identical, but different self). By treating these two cases as analogous, she makes it seem like plausibility transfers from the one case to the other. But since, as I will show, the analogy breaks down, her claims for the 'clear and positive' conception of diachronic

unity break down as well. This in turn calls into question her argument for the simple self based on diachronic unity.

As I have argued, knowing how to use 'I' in the synchronic case involves an understanding that it applies to all the co-conscious experiences in my awareness. Any experience that my present awareness encompasses is mine, and any experience that lies outside of this awareness is not mine. As William James (1980, 1. 226) points out,

> …the breeches between…thoughts…belonging to different…minds…are the most absolute breeches in nature.

To help our understanding of the diachronic case, Nida-Rümelin (2010) lays out a thought-experiment. The thought-experiment in question involves the splitting of a brain each half of which survives in a separate body. Ironically, Parfit (1984) discusses this type of thought-experiment specifically to discredit the Simple View; in contrast, Nida-Rümelin thinks that this very thought-experiment can be used to illustrate and bolster the Simple View.

The case involves surgeons dividing a brain in half (with the understanding that each of these halves are fully capable of functioning like the original) and transplanting them into two donor bodies. Nida-Rümelin asks us to imagine undergoing such a division ourselves, and see if we would not wonder which resulting person we would end up being. Indeed, there is a temptation to wonder about this, even to be convinced that there must be a determinate fact about it. But she goes a step farther than just registering this temptation; we have, she says, a *clear and positive* conception what each scenario – I end up as being the one, I end up as being the other – would involve. We can grasp the distinction by applying 'I' first to the one person and next to the other. What we here contemplate is the possibility that "I" wake up in the one body or the other after the operation. And since we grasp this distinction clearly, it would be absurd to deny that we are thinking about two different states of affair, even though they are, by assumption, indistinguishable in terms of any physical or psychological fact.

She argues that this understanding of the two scenarios is so deeply ingrained in our conceptual schemes that it is independent of philosophical theories of personal identity, so, even a person who subscribes to the psychological or physical continuity account of personal identity would have the very same conception of what it means to be the same person at some future time. Thus, even proponents of the psychological or physical continuity account of personal identity would have to, if they seriously reflect on their own subjective understanding of diachronic unity, see the difference between these two scenarios.

Nagel (1986, 33-34) explains how this works succinctly:

> My nature . . . appears to be at least conceptually independent not only of bodily continuity but also of all other subjective mental conditions, such as memory and psychological similarity . . . At the same time, it seems to be something determinate and nonconventional. That is, the question with regard to any future experience, "will it be mine or not?" seems to require a definite yes or no answer… This seems to leave us with the conclusion that being mine is an irreducible, unanalyzable characteristic of all my mental states, and that it has no essential connection with anything in the objective order or any connection among those states over time.

Nagel, despite this, ends up rejecting the Simple View. Nida-Rümelin, on the other hand, assumes that our 'clear and positive,' and, according to her, nature revealing, conception of diachronic unity is correct, and that the thought-experiment therefore reveals there are facts over and above physical and psychological facts. Most plausibly, she argues, these facts involve simple, non-physical, conscious subjects, i.e., simple selves. Denying this would mean that conceptions that we cannot help having are illusory, a bad result.

She even claims (Nida-Rümelin 2010, 239) that any alternative way of thinking about the self would be defective, would change the subject, and not really be about the self at all:

> In order to bring ourselves to seriously believe that the apparent capability to grasp what the difference consists in is illusionary we would have to lose the capacity to think about our own past and future in the first-person mode and we would have to lose the capacity to conceive of others as subjects of experience. There is reason to reject any theory that can only be seriously believed by beings that are conceptually impoverished in such a dramatical and undesirable manner.

But a person who does not hold the Simple View need not have any difficulty thinking about oneself subjectively; I don't think Reductionist views about the self are defective. What is more, I don't think that our grasp of the supposed difference between the two scenarios in the brain division case is particularly clear. This is because the concept 'I' by itself does not afford us insight into the nature of the self.

The concept 'I' indeed is independent of any empirical criteria of personal identity – like any indexical, it is simple, direct, and without descriptive content. We apply it without the mediation of any behavioral, physical, functional, or psychological criteria. Consequently, using the concept 'I', we can separate ourselves in imagination from *any* of our properties, and conceive of waking up tomorrow as a mighty elephant, a tiny mouse, or a benevolent dolphin, or, as in our thought-experiment, as one or the other person after the brain-division surgery. This is what grasping what it means to be one or the other future person comes down to: that we can apply the concept 'I' to either of them in imagination. But exactly because this, that we can apply it in imagination to *any* person, past or future, it doesn't facilitate a 'clear and positive' grasp of what it means to be the same person in the future.

Of course, when I think about my future self, e.g., hoping that I will finish this paper next week, I use the concept 'I' in a direct, simple, and straightforward way. It does not lead to confusion or uncertainty. It succeeds in latching on to its referent because certain constitutive conditions in my psychological and physical make-up – including undivided memory continuity – are being met. But merely being able to deploy the concept 'I' does not give me a clear idea of what it is to be a persisting self over time.

In claiming that we have a 'clear and positive' conception of the division scenarios, Nida-Rümelin might be conflating the synchronic and diachronic deployments of 'I'. In the synchronic case, I have a clear and positive insight into what it means to be me at the present moment – even though this insight does not include anything about a simple self. So, it seems initially plausible that I also have a clear and positive understanding of what it takes to be me as opposed to be someone else just like me in the future. But the basis of my understanding in the synchronic case is not present in the

diachronic case. In the synchronic case, my understanding of what makes an experience mine at this moment is based on reflexive self-awareness; contrarily, my understanding of what makes an experience mine in the future is, by definition, *not* based on reflexive self-awareness. One way to see that the synchronic case is radically different from the diachronic case is that whereas in the synchronic case I can tell myself apart from my twin with absolute certainty, there is no way for me to know which of the future selves (if any) in the division case is me; and neither one of my successors can know this either (though both will be inclined to say they are the one).[16] We use the same indexical concept 'I' in both cases, but its use in the synchronic context is different from its use in the diachronic one.

And because of that, when it comes to imaginary cases that deviate in crucial ways from the usual contexts in which 'I' is normally used, my grasp on the situation diminishes. In the division case, I strain to understand what the difference in my relationship to the two resulting selves could possibly be. Even worse, the same criterion-free use of the 'I'-concept makes it look like certain scenarios are conceivable which are clearly inconceivable – such as having been Ludwig Wittgenstein or going to become an animal.

This suggests that our first-person conception of diachronic identity is impoverished and confused, rather than positive and clear. The illusion of clarity comes about because of an image of sorts. I project into the future the idea of the inside view of what it means to be in *my* 'head' at this moment, as opposed to a qualitatively identical 'head': I imagine having an inside view of a tunnel along which my conscious states flow, and which exclude all other conscious streams no matter how similar. Of course, no such tunnels exist. Despite Nida-Rümelin's insistence to the contrary, we simply do not have a substantive grasp of what it would be to be in *my* 'head' in the future as opposed to a qualitatively identical, but distinct other 'head'. Here is Korsgaard (1989, p. 116) on this:

> [consciousness]…is envisioned as a tunnel or a stream, because we think that one moment of consciousness is somehow directly continuous with others, even when interrupted by deep sleep or anesthesia… The sense that consciousness is in these ways unified supports the idea that consciousness requires a persisting psychological subject.

The tunnel metaphor in fact does not *support* the intuition that there is a difference in outcome in the division case depending on which new person is me; rather it *presupposes* it. We do not have a clear and positive conception of the two scenarios (a clear and positive grasp of these 'tunnels') that somehow reveals the existence of simple selves. Rather, it is holding the Simple View that allows us to have a 'clear and positive' conception of the difference between the two scenarios, and so to think about these two streams as either mine or not mine. Consequently, this allegedly clear and positive conception of the case cannot in turn be used to justify the Simple View: the conception of the difference itself is an upshot of the Simple View. Our uncertainty in the division case as to what it would mean for us to be the one or the other future self gives way to a 'positive' understanding only after we suppose that the Simple View is true.

The Simple View is certainly intuitive. But it is not part of how any normal human thinks of diachronic unity; it is a particular conception of it. Because of the bareness of the criterion-free conception of the self Nida-Rümelin advocates – solely based on how the concept 'I' is used in

[16] For the same reason, the Simple View then opens up a gratuitous skeptical worry about whether I am really the same person from day to day – since according to it, no empirical relation such as memory could settle the issue.

thought – her conception conflicts with some other common conceptions of the self. On her view, memory continuity is not constitutive of personal identity, and can be entirely broken without personal identity being broken. This agrees with the sensibilities of those who believe in reincarnation but goes against the more widespread belief that one's survival is not conceivable in the absence of any memories. The notion that it is conceivable for me to have been Wittgenstein seems to me to be wrong, but such a scenario is not ruled out by the 'I'-concept alone.

The Simple View goes astray as it is a reification of our 'I'-concept. It takes features of the concept – its simplicity and independence from empirical relations – and attributes these features to the referent. But our concept of the self is not purely subjective; the subjective conception would only give us the momentary self. The self-concept involves more than reflective self-awareness in the present moment, or direct deployments of 'I' across time. It also has an objective component singling out what lends unity to the direct deployments of 'I' over time – and it is not the simple self but memory, and other psychological continuity. To form a full-fledged concept of the self – one's own, and other people's – existing across time, one needs to apprehend objective criteria for the identity of selves, in addition to being able to deploy 'I', from the inside, to past, present or future experience. As Jenann Ismael (2007, 176) explains:

> the process by which the purely mechanical procedure of attaching indexicals to thoughts is transformed into real, identifying thought – for example, the process by which one goes from merely producing mental 'I'-tokens to actually *thinking about oneself* – then, involves grasping the truth of an indexical identity sentence. To represent oneself in the sense of being able to take oneself as an *object* of thought requires more than the ability to produce mental tokens of 'I'. It involves grasping the objective truth conditions that integrate those tokens into an articulated network of concepts.

In sum, the argument from diachronic unity does not go through, as we lack a clear and distinct conception of the alternative scenarios in the division case, just as Parfit argues. The alternative scenarios seem to be possible given our ability to freely deploy the 'I'-concept from the subjective point of view across time. But at the same time, they also seem impossible given our objective understanding of cross-temporal identity as constituted by memory relations. The right response to the division cases is puzzlement, not clarity. This does not mean that the Simple View is false. But it means that these kinds of consideration do not give reason to believe it, while the alternative, Reductionist account remains fully viable.

As for Nida-Rümelin's complaint that if the Simple View were false, we would be victim of a persistent illusion that we cannot help having, we should remember that such persistent illusions are not unusual. For example, people tend to think that objects are solid through and through even after it is pointed out to them not to be the case. Or, more pertinent to the case, many people find the libertarian conception of free will compelling. According to this view, often held by proponents of the Simple View, we are capable of action that is independent of everything that happened in the universe until then. Such views are incompatible with science. Libertarian free will, in particular, is in conflict with the causal completeness of physics – the thesis that every physical event has a complete explanation in physical terms – which has been particularly well corroborated in the past century.[17] One reasonable response to this situation is to give up the old conception of free will and introduce

[17] For a history of how this thesis became widely accepted by both scientists and philosophers, see Papineau (2001).

a new concept more in line with what we know about the world. One could do the same regarding the Simple View.

3. The Practical Significance of Metaphysics

Beliefs about the metaphysics of the self are not only of theoretical interest; metaphysical beliefs also have practical significance. For example, Buddhism considers an understanding of the insubstantial, impermanent nature of the self essential to ending suffering. The opposing, Simple View – like the one Nida-Rümelin argues for – is thought to be not only false; it is considered practically harmful as it fuels exaggerated concern for oneself and closes one off from sources of joy and meaning. Coming to believe that the simple self does not exists has a liberating and edifying effect, they claim.

To the contrary, Nida-Rümelin thinks that the concept of the simple self is so deeply embedded in our thinking about moral and prudential significance, that any ongoing commitment to others or to oneself is only appropriate only if it exists. Another way to put it is that Reductionism would render our attachments, our love and care for one another illusory (Nida-Rümelin 2010, 240):

> Once we realize that the criterion-free notion of identity across time for conscious individuals is present not only in our thinking but also in our perceptions and emotions it becomes clear that the illusion at issue is still more general and deeper than one might think at first sight. When a person is touched by meeting a friend that she has not seen since many years then she perceives that friend as identical to the younger person she knew so well in the distant past. Perceiving the other person in that way which incorporates the criterion-free notion of identity is an essential component of that emotional experience. Following this line of thought it becomes clear that most of what we value in life would be based on a fundamental cognitive illusion if the illusion theorist was right.

But if the *truth* of Reductionism would undermine the reason to care about our future or the lives of others, a *belief* in Reductionism would render our attachments simply irrational (though it is not clear if she thinks the Reductionist view can even be held in good faith). That would certainly be a rather shocking implication of Reductionism. Of course, the mere fact that a certain belief has undesirable consequences is no argument against the *truth* of that belief. I have addressed Nida-Rümelin's point about persistent illusions in the previous section, purely from the point of view of a theoretical interest in the truth. But the point about the practical consequences of metaphysics cannot be dismissed simply on these grounds; the relationship between metaphysics and matters of value is a delicate and important issue. In the rest of the paper, I want to consider various aspects of the relationship between Reductionism and the practical concerns of our lives.

One such connection can be quickly dismissed. The claim that if we stopped believing in the simple self, we would also, as a matter of empirical fact, would stop caring ourselves and one another, seems to be lacks empirical support; there is no reason to suppose that this would generally be the case. Humans are hard wired to care about what happens to them next; they are also hard wired to care about – at least certain – others. This is a fact of evolution. Nida-Rümelin might argue that this is due entirely to our innate tendency to grasp what it is to be a self in terms of the simple self. However, the existence of Buddhist thought refutes this. One of its fundamental tenets is that the simple self does not exist, but Buddhists do not stop caring as a result.

A more interesting idea is that it is *irrational* to care if the Simple View is false. Nida-Rümelin seems to argue that the *reason* for caring somehow presupposes that people have a simple soul; and short of such a reason our cares would become irrational. I think this is not the case either. The truth of Reductionism, or even a belief in Reductionism doesn't render love of family and friends irrational, as I will argue in the next section.

On the other hand, even if the connection between metaphysical belief and evaluative attitude is not a *rational* one, there are rich and complicated *psychological* connections between them. For example, looking at everyone as a creature of God might make it easier for some people to feel empathy for others. And ceasing to believe in the simple self – even if it does not end all concern for the future, might subtly alter one's relationship to the future. In this last section, I follow a two-prong approach: in the first part, I argue that Reductionism doesn't render our cares and attachments irrational. In the second part, I concede the point – developed in great detail in the Buddhist literature, but supported by observations about other belief systems about the self as well – that one's metaphysics might, as a matter of psychological fact, influence one's practical attitudes in subtle and sometimes obvious ways.

3.1 Does the Value of Personal Identity Depend on the Simple View?

The claim I aim to refute, which Parfit (1984, Chapter 14, 307) calls the "Extreme Claim", is that on the Reductionist account, there is no reason whatsoever to care about personal identity. Nida-Rümelin is not alone in her grim assessment of what the falsity of the Simple View would imply for our lives. In Sidgwick's (1907, 418-9) words:

> Why… should one part of the series of feeling…. be more concerned with another part of the same series, any more than with any other series.

Or, as Swinburne's (1973-4, 246) puts it, if all there is to personal identity is certain continuities in experience, "in itself such continuity has no value".

Parfit puts forward (1984, Chapter 14, 311) what he calls the "Moderate Claim": that certain empirical relations, such as memory continuity and connectedness – which in normal cases, i.e., cases not involving division, duplication, etc. underlie personal identity – provide a reason for special egocentric concern for one's own future. The same holds for concern about significant others; lovers, friends, relatives. The fact that they bear these relations to their past and future selves, provide reason for our continued care about them.

I agree with the Moderate Claim. I think there are no simple subjects. It is fundamental to human life that that we love and cherish others which involves both appreciating their past and caring for their future. What proponents of the Simple View assert, and Reductionists deny, is that the reason for such care lies in simple selves. But the proponent of the Simple view is on thin ice when they predicate value on the existence of simple selves, *whether or not* such selves exist. The Reductionist, of course, by rejecting the Simple View, will thereby also reject the idea that the simple self is the locus of value.

It is one thing to argue that the Simple View is true and that given what we are – simple selves – it is simple selves that we care about when we care about personal identity. But it is another thing

altogether to claim that *even if this view were false,* we still would only have reason to care about simple selves and would have no reason to care about what *in fact* constitutes personal identity over time, namely, memory – and other psychological – connections. To establish such an implausible claim needs a strong argument. Nida-Rümelin's argument is that the very concept of self analytically implies that it is a simple, non-material entity. If this were so, believing that there are no simple selves would be tantamount to believing that there are no persons at all; so nothing to care about.

As I argued in previous sections, however, our concept of self does not by itself imply simplicity even if the idea of a simple self comes natural to humans. In fact, everything that matters in caring about other people's or our own future can be described in terms that are neutral with respect to the metaphysical truth about selves; 'self' or 'person' can be used to fully describe and delineate one's attachments and cares about the social world without ever implying anything about its nature.

I would venture that one need not have metaphysical views about the self at all to live a normal human life. Nida-Rümelin's claims to the contrary seem to me an exaggeration. Sure, some people do hold the Simple View and perhaps this view is closely intertwined with their attitudes about personal identity and the future. But others do not, and, as the example of Buddhism shows, whole cultures might be built on an altogether different notion of the self. In other words, from a Reductionist perspective, the value of personal identity doesn't – and cannot – depend on the existence of the simple self.

And, even worse for the Simple View, identity of the simple self *in itself* – even if such selves existed – would not be enough. If such identity were preserved but without any continuity, caring for that person's future would largely lose its point. Should you become an elephant tomorrow, inhabited by the same simple self as before but having no memory of your life, or even human-like mentality, I would not have the same reason to care about you as I had before. Speaking for myself, I would not find it consoling to survive death if this survival would erase all my memories; on the other hand, and I would be happy to know there is someone continuing my explorations, carrying my memories, living my feelings etc. When it comes to the fundamental locus of the value of personal identity, memory connectedness and continuity will do.

3.2 The Practical Significance of Metaphysical Beliefs

But it is one question whether it is irrational to care about the future on the Reductionist view, and another question whether believing Reductionism or the Simple View has different effects on our practical attitudes. While I believe, as I argued above, that people can live normal lives and have normal attachments without any metaphysical views at all, I find it obvious that beliefs of a metaphysical sort about, say, the existence of an immortal soul, or the presence of spirit in everything do, as a matter of psychological fact, have consequences about what we value, how we live. These beliefs might even be the main vehicles in which a culture transmits values. Metaphysical beliefs shape what matters in the way like mythology or religion does, by affecting and organizing our experience. These beliefs encompass how we view our place in the world in the most general terms.

Our evaluative outlook depends on how we experience the world. One of the main vehicles to becoming aware of the value of things is contemplating our experience (Balog 2020). So it is via intermingling with experience that metaphysical beliefs make a practical difference. In a quite literal sense, our metaphysical beliefs about the self concern not only what we *think* about ourselves, but they have an effect on what we are like what we value, etc.[18] An experience of an awesome presence will be modulated and colored by the particular religious beliefs one has or doesn't have. An atheist or Buddhist will have a different experience of that presence from a religious person. There is much to say about this topic which I won't be able to do in this paper. I will limit myself to some broad observations about the Simple View on the one hand and Buddhist Reductionism on the other, to illustrate the kinds of questions involved.

According to Buddhism, although we have a natural affinity for the Simple View, we are well-advised to transcend it. The Buddha ties the adoption of the notion of the simple, substantial self to our desire to pacify the terror we feel when we acknowledge the inevitability of death. According to Buddhism, this view involves us in a fruitless search for the simple self in our experience, and results in a kind of self-centeredness that stands in the way of happiness and ease. Our experience is shot through with a sense of ourselves as solid and unchanging which prompts exaggerated concerns about ourselves. We are forever engaged in the project of trying to become a kind of self who wouldn't have undesirable experiences.[19] According to Buddhism, the real profundity of experience can only be revealed when the anxious effort to protect the self is given up, when the impermanent, unsubstantial nature of the self is acknowledged. This is what the Buddha calls enlightenment.

But the Simple View has its attractions. It affirms a sense that our lives flow from past to future in a single channel. Reductionism would have us accept that it is not necessarily so and that carries an air of groundlessness. The Simple View provides a bulwark against that groundlessness. It provides an orderly image of the self, a counterweight against the chaotic impermanence of our lives. Its most common form is a belief in the existence of immortal souls, which opens up the possibility an afterlife, and a meaningful order in the world.

Another offshoot of the Simple View, the concept of the self as free agent, undetermined by nature rose to prominence during the Enlightenment. Charles Taylor (1989, 364), describes this conception as

> a radical definition of freedom, which rebels against nature as what is merely given, and demands that we find freedom in a life whose normative shape is somehow generated by rational activity . . . a powerful, it is not overstated to say revolutionary, force in modern civilization. It seems to offer a prospect of pure self-activity, where my action is determined not by the merely given, the facts of nature (including inner nature), but ultimately by my own agency as a formulator of rational law.

Compared with this conception of radically free and self-created individuals, the Reductionist view, by contrast, can be demoralizing. As Bernard Williams points it out in an interview[20]: "we are, as

[18] For the "looping effect" that characterizes the interaction between how we think about ourselves and others and the selves so classified, see Hacking (1986).
[19] See Kierkegaard (1849/1982) who makes the same point.
[20] Interview with Bernard Williams at https://manwithoutqualities.com/2015/03/28/interview-with-bernard-williams-2/.

Nietzsche said in one of his many images on this subject, a kind of polyp. Out of this mass of stuff emerge our actions."

A recent study on the effects of psychedelics on metaphysical views and wellbeing (Timmermann 2021) suggests the Simple View is associated with, and might even contribute, to a sense of meaning and wellbeing in the world. It is possible that the Reductionist understanding of the self, despite Buddhist claims to the contrary, is not as conducive to human flourishing.

While I think Nida-Rümelin is wrong about the Extreme Claim, that is, the claim that on a Reductionist account, we have no reason to care, a case might be made that the Reductionist understanding of the self represents a loss in our relationship with ourselves and others. It is an open question if this is so, and if so, how this loss can be overcome.

Bibliography:

Balog, Katalin (2012). Acquaintance and the Mind-Body Problem. In Christopher Hill and Simone Gozzano (Eds.), *New Perspectives on Type Identity: The Mental and the Physical* (pp. 16-43). Cambridge: Cambridge University Press.

Balog, K. (2020.). Either/Or: Subjectivity, Objectivity and Value. In E. Lambert, & J. Schwenkler (Eds.), *Becoming Someone New: Essays on Transformative Experience, Choice and Change,* 254-69, Oxford University Press.

Bering, J.M. (2006). The folk psychology of souls. *Behavioral and Brain Sciences, 29*, 453-498.

Brentano, Franz (1874/1973). *Psychology from an empirical standpoint,* Routlede & Kegan Paul.

Butler, Joseph (1736/1896). *The Analogy of Religion*, Oxford: Oxford University Press.

Chalmers, D. (1996). *The Conscious Mind.* New York : Oxford University Press.

Chisholm, Roderick (1976). *Person and Object.* Dordrecht: D. Reidel.

Descartes, René (1985). *The Philosophical Writings of Descartes,* vols. 1 and 2, trans. J. Cottingham et al., Cambridge: Cambridge University Press.

Gertler, Brie (2012). "Conscious states as objects of awareness: on Uriah Kriegel, Subjective consciousness: a self-representational theory.", Philosophical Studies 159: 447-455.

Hacking, I. (1986). Making Up People. In D. Sosna, M. Wellbery, & T. Heller (Eds.), *Reconstructing Individualism* (pp. 222-36). Stanford, CA: Stanford University Press.

Hume, David (1739/2000). *A Treatise of Human Nature*, edited by David Fate Norton and Mary J. Norton, Oxford/New York: Oxford University Press.

Ismael, Jenann (2007). *The Situated Self*, Oxford University Press.

James, William (1890/1950). *The Principles of Psychology*, 2 vols., New York: Dover.

Kant, Immanuel (1781-87/1998). *Critique of Pure Reason*. Cambridge University Press.

Kierkegaard, S. (1849/1982). *The Sickness Unto Death*. Princeton University Press.

Kriegel, Uriah (2009). *Subjective Consciousness: a Self--Representational Theory*, Oxford University Press.

Korsgaard, Christine. 1989. Personal identity and the unity of agency: A Kantian response to Parfit. *Philosophy and Public Affairs* 18, no. 2: 101-132.

Locke, J. (1689/1996). *An Essay Concerning Human Understanding*. Indianapolis/Cambridge: Hackett Publishing Company.

Lycan, William (2004). "The superiority of HOT to HOP", In R. Gennaro (ed.), Higher Order Theories of Consciousness: an Anthology. Philadelphia: Benjaminis, 93-114.

Nagel, T., *The View from Nowhere*. Oxford: Oxford University Press, 1986.

Nida-Rümelin, Martine (2010). "An Argument from Transtemporal Identity for Subject Body Dualism", in George Bealer and Robert Koons (eds.), *The Waning of Materialism*, Oxford University Press.

--- (2011). "The Conceptual Origin of Subject Body Dualism", in Annalisa Colliva (ed.), *Self and Self-Knowledge*, Oxford University Press.

--- (2017a). "Realism about Identity and Individuality of Conscious Beings", in Christian Kanzian, Sebastian Kletzl, Josef Mitterer, Katharina Neges (eds.), *Realism - Relativism - Constructivism: Proceedings of the 38th International Wittgenstein Symposium in Kirchberg*, Berlin, Boston: De Gruyter, 279–292.

--- (2017b). "Self-Awareness", *Review of Philosophy and Psychology* 8 (1): 55-82.

O'Connor, Timothy (2000). *Persons and Causes,* Oxford University Press.

Papineau, D. (2001). The Rise of Physicalism. In B. L. Carl Gillett (Ed.), *Physicalism and its Discontents* (pp. 3-37). Cambridge: Cambridge University Press.

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

Rosenthal, David (2002). "Explaining consciousness", in D. Chalmers (ed.), Philosophy of Mind. Classical and contemporary readings.", Oxford University Press, 109-131.

Reid, Thomas (1785/2011). Essays on the Intellectual Powers of Man. Essay Three: Of Memory, Chapter 4: Of Identity.

Sartre, Jean-Paul (1943/1956). *Being and Nothingness*, New York: Philosophical Library.

Siewert, Charles (2013). "Phenomenality and Self-Consciousness," *Phenomenal Intentionality*, edited by Uriah Kriegel, Oxford University Press.

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

Strawson, Peter F. (1959). *Individuals*, London: Methuen.

Swinburne, Richard (2007). "From Mental/Physical Identity to Substance Dualism", in P. Inwagen and D. Zimmerman (eds.), *Persons: Human and Divine*, Oxford University Press.

Taylor, C. (1989). *Sources of the Self: The Making of the Modern Identity*. Cambridge: Cambridge University Press.

Thompson, Evan (2011). "Self-No-Self? Memory and Reflexive Awareness", in: *Self, No Self?: Perspectives from Analytical, Phenomenological, and Indian Traditions* (eds. Mark Siderits, Evan Thompson, Dan Zahavi), Oxford University Press, 157-176.

Timmermann, C. *et. al.* (2021). Psychedelics alter metaphysical beliefs, *Nature 11:22166,* (Bering, 2006).

Velleman, David (2015). "So It Goes", in *Beyond Price: Essays on Life and Death*. Cambridge, UK: Open Book Publishers, http://dx.doi.org/10.11647/OBP.0061

Zahavi, Daniel (2006). "Thinking about self-consciousness: phenomenological perspectives.", in U. Kriegel and A. Williford (eds.), Self-representational Approaches to Consciousness, Cambridge MA: MIT Press, 273-295.