Fission, First Person Thought, and Subject-body Dualism[1]

Kirk Ludwig
Philosophy Department
Indiana University

1. Introduction

Imagine perfect fissioning of a person A into two successors who are equally good candidates for being the person who underwent the fissioning, LA (on the left) and RA (on the right), on the basis of all the material and psychological relations (empirical relations for short) that each bears to A. Imagine that if LA had been produced by the process without RA, we would judge that A = LA, and vice versa. We seem, *prima facie*, to understand what it would be for A to survive as (be identical with) LA (and not RA), and what it would be for A to survive as (be identical with) RA (and not LA). Does this show that

[SBD] A is not identical with, or constituted by, or composed (even in part) by a body, or any material stuff, or even any immaterial stuff.

Call [SBD] the thesis of subject-body dualism.

In (Nida-Rümelin 2006) and (Nida-Rümelin 2010), Martine Nida-Rümelin (NR) argued that these considerations do establish [SBD]. The master argument has three premises:

[1] My thanks to helpful conversation with Martine Nida-Rümelin and comments from two referees for this journal.

1. There is a *factual difference* between the claim that someone is one or the other of the two continuers in fission cases or we are subject to a pervasive illusion in our thoughts about personal identity over time (call this *the illusion theory*).
2. There could be a factual difference between the claims that someone is one or the other of the two continuers in fission cases only if subject-body dualism were true.
3. The illusion theory is untenable (false).
4. Subject-body dualism is true [1-3].

Granting premise 3 for the sake of argument, we get the streamlined argument (when I refer to premises, I will have the streamlined argument in mind):

1. There is a factual difference between the claim that someone is one or the other of the two continuers in fission cases.
2. There could be a factual difference between the claims that someone is one or the other of the two continuers in fission cases only if subject-body dualism were true.
3. Subject-body dualism is true [1-2].

The basic idea of the argument is that the materialist can't make sense of there being a factual difference between the claim that someone is one or the other of two continuers in fission cases. The reason is that the materialist has to appeal to bodily or psychological continuity or some combination of these to ground claims about transtemporal identity (call these empirical criteria of transtemporal identity), but in the fission case there is complete symmetry with respect to all the empirical criteria between A and RA, on the one

hand, and A and LA, on the other. Whatever you can say about either case you can say equally about the other. How then can there be a factual difference between the two claims?

In (Ludwig 2013), I argued that on the three most plausible interpretations of 'factual difference' in premise 2, the argument failed. Consider our initial case, and let 'D' describe the relevant facts completely except for the facts about identity between A, LA, and RA. Then consider the claims P1-P3.

P1: D and A = LA.

P2: D and A = RA.

P3: D and A ≠ LA and A ≠ RA.

We assume that LA ≠ RA. In terms of this case, the three interpretations of 'there is a factual difference between P1 and P2' are these:

(1) P1 is true and P2 false *or* P1 is false and P2 is true.

(2) P1 and P2 differ in content (express different propositions).

(3) P1 and P2 both express genuine possibilities.

(1) says that there being a factual difference is there being a fact of the matter which is true and which false. (2) says that the factual difference is a matter of P1 and P2 expressing different propositions. (3) says that it is a matter of their expressing genuine possibilities (they might express different propositions but neither express something that is possible).

Briefly, I argued in (Ludwig 2013) that (1) cannot be the right interpretation because NR allows that P3 is a genuine possibility and she doesn't argue it is not actual. I argued that (2) isn't sufficient for the argument because the materialist can make sense of there being a difference in content between P1 and P2 by saying that that 'A = LA/RA' means 'LA/RA's body is A's body'. Finally, I argued (in part—I return to this below) that (3) isn't sufficient because, even granting that, in a world in which P1 or P2, A would not have empirical criteria of transtemporal identity, it would be a modal fallacy to infer we don't *actually* have empirical criteria of transtemporal identity.

NR responded in (Nida-Rümelin 2013) that (i) the criticism fails because it overlooks the intended interpretation of factual difference (let's call this the fourth interpretation, which I will explain below), and that (ii), in any case, the third interpretation is sufficient for the argument to go through, when supplemented with the assumption that objects have their transtemporal identity conditions essentially. I respond in this paper.

In section 2, I take up the fourth interpretation, identify the underlying assumption that motivates it (that in first person thought we have a positive conception of the self that excludes empirical criteria of transtemporal identity), and formulate an argument for subject-body dualism that rests directly on that assumption. In section 3, I argue that we do not have good reason to accept the assumption because relevant features of first person thought are accommodated by our referring to ourselves directly and not under any mode of presentation (or by description). In section 4, I respond to the claim that the argument on the third interpretation is successful when combined with the assumption that we have our criteria for transtemporal identity essentially. I argue that the third interpretation

requires a contradiction be true because it is inconsistent with the necessity of identity, so that the only possibilities we could be thinking of are epistemic possibilities. I explain this in terms of the account of first person thought developed in section 3. Section 5 is a short conclusion.

2. The Fourth Interpretation and the Reformulated Argument

The fourth interpretation is expressed in the following passage:

[4I] … the fourth (and originally intended) interpretation of the claim that there is a factual difference [between] P1 and P2 … can be made explicit in the following way: *there is an objective feature which would have to be realized* ***in addition*** *to what description D states for P1 to be realized, and there is (a different) objective feature which would have to be realized* ***in addition*** *to what description D states for P2 to be realized.* (Nida-Rümelin 2013, 705)

NR says that (i) "that there is a factual difference between P1 and P2 is compatible with the metaphysical *impossibility* of both P1 and P2," for (ii) it "does *not* imply that the fulfillment of description D is metaphysically compatible with the relevant feature which would render P1 actual" and likewise for P2, and "so it does not imply the metaphysical possibility of P1 or the metaphysical possibility of P2" (loc. cit.).[2] As the last clause indicates, (ii) is the ground for (i).

[2] In her (Nida-Rümelin 2010), NR wrote: "This difference appears to be factual in this sense: 'D and Andrea is L-Andrea' and 'D and Andrea is R-Andrea' are not just two legitimate description[s] of one and the same course of events. Rather there is—according to the way we conceive of the situation—an objective *possible*

Would a factual difference in this sense help us to understand why premise 2 of the (streamlined) argument should be accepted?

I do not think that [4I] is sufficient, but I also think that there is more in the background than is expressed here, and we will come to this in a moment. The reason that [4I] is not sufficient is that a proponent of empirical criteria of personal identity over time could accept it as it stands. If one thinks that our criteria for personal identity over time involve bodily or psychological continuity (or both), then in fission cases, given that LA ≠ RA, that A is not identical with the pair of LA and RA, and that the symmetry of the case precludes choosing one over the other, the only option is to say that A does not survive. A does not survive because survival requires there be a unique closest continuer of A (which is close enough—let this be understood). Where there is no unique closest continuer, A does not survive. (This is what we say about cell fission.) How can the proponent of an empirical criterion of personal identity accept that *there is an objective feature which has to be realized in addition to what D states for P1 to be realized*? Given that it does not have to be a feature which is compatible with D, she can say that the proposition that LA is *the best* continuer of A expresses the relevant objective fact that you would have to add—though it turns out to be incompatible with D, which entails that there is no best continuer of A. Since it is incompatible with D, it is clearly something in addition to what D expresses that

---

feature of the world *that makes one of the two descriptions true and the other wrong*. The factual difference may be described [by] pointing out that Andrea will have a different future depending on which of the two *possible* identity facts will obtain" (2010, 196). One might be forgiven for thinking NR was assuming that P1 and P2 were possible, and that the relevant possible feature of the world, as it would make true one of the two descriptions (which here are explicitly P1 and P2), was compatible with D.

would have to obtain. If we were to add that the additional objective features were compatible with D, that is P1 and P2 were each possible, then this response would be closed off. But then this would be equivalent to the third proposal after all. As we will see in section 4, however, the third proposal is incoherent.[3]

---

[3] A referee for this journal suggested that one might say that the additional objective features (call them F1 and F2 respectively) were each compatible with D and that this was still compatible with P1 and P2 being metaphysically impossible. So the idea is this: that F1 obtains and D is possible, and P1 entails that F1 obtains but not vice versa, and that F2 obtains and D is possible and P2 entails that F2 obtains, but not vice versa, and that is what the factual difference comes to. But it is not required that P1 or P2 be possible. However, this divides into two cases. Either P1 or P2 is possible or neither P1 nor P2 is possible. The first case collapses into proposal three (see note 11). So we may restrict attention to the viability of the claim that F1 and F2 are both compatible with D but P1 and P2 are impossible.

This is not NR's *intended* interpretation. On the intended interpretation, the factual difference "does *not* imply that the fulfillment of description D is metaphysically compatible with the relevant feature which would render P1 actual" and likewise for P2. Therefore, as noted, on NR's interpretation of the factual difference, it could involve empirical criteria of transtemporal identity, such as being the closest (close enough) bodily continuer, since the features don't have to be compatible with D.

Still, could NR appeal to this new suggestion instead? Not without rejecting the reasoning she actually engages in. I noted that (ii) is supposed to be the reason for (i).

(i) the factual difference is compatible with the impossibility of P1 and P2,

(ii) the factual difference "does *not* imply that the fulfillment of description D is metaphysically compatible with the relevant feature which would render P1 actual" and likewise for P2.

NR states (ii) and then immediately writes: "so it [the factual difference] does not imply the metaphysical possibility of P1 or the metaphysical possibility of P2." This is why she says that the factual difference is compatible with the impossibility of P1 and P2, that is, (i). More specifically, (ii) is the ground for (i) because the idea is that if only F1 could be compatibly added to D that would suffice for A = LA, and similarly for F2

and A = RA. That is why we get to say that F1/F2 are objective features that would be just what has to be added to D for P1/P2 to be realized. But if it is left open that they are not compatible with D, it is left open that P1 and P2 are not possible. If we said instead that D was compatible with F1, then, by hypothesis, the relation required for A = LA would be realized in some possible world, and so A = LA is possible. And if we said that D was compatible with F2, then the relation required for A = RA would be realized in some possible world and A = RA is possible. But if either is, surely the other is, given the symmetry of the situation, and so we would be back with interpretation 3.

Why not reject the reasoning? Couldn't one just *insist* that F1 is compatible with D (mutatis mutandis for F2) and that P1 and P2 are impossible? But now why are P1 and P2 impossible? It is not because of the symmetry involving empirical criteria of transtemporal identity, because we are rejecting empirical criteria of transtemporal identity. So if the objective non-empirical feature F1 that would be just what has to be added to D to realize P1 is compatible with D, there is a possible world in which D and the objective feature F1 which we need to add to D to realize P1 are co-realized. That is all then that is needed, given what we have said, to realize P1. So P1 is possible. Otherwise there is no content to the idea that the feature F1 is an objective feature that is what is to be added to D to realize P1. Similarly for F2 and P2.

What if one said: but why do F1 and F2 have to be sufficient for P1 and P2 (given D)? Why can't they just be *necessary*? NR aims to show that transtemporal identity of subjects of experience is not grounded by empirical criteria. Thus, in a case like the fission case except that LA/RA was the only survivor, there would be an objective feature, F1+/F2+, that was non-empirical which in the circumstances sufficed for A=LA/A=RA. Thus, F1+/F2+ is the feature that would have to be realized in addition to D for P1/P2 to be realized, where it is not merely a necessary condition but sufficient, given that D duplicates whatever empirical relations there are between A and LA/RA, in the circumstances in which F1+/F2+ are sufficient for A=LA/A=RA. And now we are back to where we started. So the trouble is this: if we want to acknowledge that P1 and P2 are both impossible, but we want to reject empirical criteria of transtemporal identity, we can't say the non-empirical condition that is all that needs to be added to underwrite P1 (mutatis mutandis for P2) is compatible with D.

However, [4I] does not fully express the underlying thought. We need to look at the ground for the claim that there is an additional objective feature that would have to be added. NR says that we

> [a] … have a *clear positive understanding* of an objective feature of the world that would make it the case that one of the two successors (and not the other) is the original person. [i] If, for instance, P1 is realized, then the original person Andrea [my A] is living L-Andrea's life after the operation. [ii] Andrea has a completely different life after the operation (she has a fundamentally different future when considered from her perspective at the earlier moment) if P2 is realized. We have the conceptual capacity to positively conceive this difference … . [b] The basic idea can be put quite simply: we understand the difference between P1 and P2 by considering it from Andrea's perspective. Thinking about Andrea *in the first person mode* we are able to grasp the difference between a world in which *she* (rigid designation of Andrea) will undergo experiences related to the body with the left brain hemisphere, and a world in which *she* will undergo the experiences related to the body with the right brain hemisphere. (703-4; labeling in square brackets added)

[b] is more fundamental here than [a]. The conditionals [i] and [ii] would usually be treated as presupposing that the antecedents are possible (so that they are not just trivially true because of a necessarily false antecedent). But we are not supposed to assume that P1 and P2 are possible because D and the objective feature that would make it the case that,

e.g., in the case of P1, A = LA was true may not be compatible. But given this, so far the proponent of the empirical criterion of identity can accept [i] and [ii].

Turn to [b] then. What is it to think about Andrea *in the first person mode*? To elucidate this, we should return to the original article (Nida-Rümelin 2010, sec. 4-6). This will reveal that what is crucial is the content of the positive conception of the difference between P1 and P2 (at least so far as what it excludes), and we will here identify, I believe, the fulcrum of the argument.

The key idea is brought out by first imagining oneself in A's position. You may wonder whether you will survive fissioning and whether, if you do, you will be PL (the person associated with the left hand body) or PR (the person associated with the right hand body). You wonder whether in the morning, if you wake up, you will see PL's face in the mirror or PR's face in the mirror, whether you will feel pain if PL's body is pricked, or if PR's body is pricked, and so on. You have a clear grasp, it seems, of what it would be to be one or the other of PL or PR from the first person point of view, that is, of what facts would be appropriate for the one hypothesis or the other from that point of view, where these facts can be stated in the form "I will have property *P* at moment *m*'. From this, NR concludes that

> Claim: 1: Transtemporal self-attributions are conceptually prior to self-identifications. (Nida-Rümelin 2010, 203)

The argument is given in the following passage:

> You understand the assumption 'I will be the L-person' on the basis of understanding thoughts like 'I will have property *P*.' In other words and more precisely: you understand what has to be the case for your utterance of 'I will be the L-person' to be true on the basis of your understanding of what would render your self-attribution 'I will have property *P* in the future moment *m*' true. We can formulate this claim in a more abstract way: transtemporal self-attributions (thoughts that can be expressed by sentences like 'I will have property *P*' or 'I had property *P*') are conceptually prior to self-identifications (thoughts that can be expressed by sentences of the form 'I will be *P* at moment *m*' or 'I was *P* at moment *m*'). (Nida-Rümelin 2010, 198)

It is important here also that this be a "clear *positive* understanding of what would have to be the case for [one's] thought 'I will be the L-person' to be true … due to [one's] clear *positive* understanding of what has to be the case for certain I-thoughts of the form 'I will have property P' to be true" (Nida-Rümelin 2010, 201, emphasis added).

NR adds to this a second, important claim, namely, that your "understanding of what has to be the case for your I-thought 'I will have property P' to be true in no way depends on the empirical criteria of transtemporal identity of subjects of experience that you implicitly accept" (Nida-Rümelin 2010, 199). NR argues for this by asking us to consider whether one's understanding of what it would be for one to be in pain in the future, for example, would change when one's explicit conception of the requirements for personal identity change. For example, suppose one confidently expects that one will survive when one's body and brain are destroyed provided that a new brain is created from a scan of the

old with the same brain states, preserving psychological continuity. One thinks, of a moment *m* after the expected destruction of one's body, 'I will be in pain at *m*'. Then suppose one comes to reject the psychological continuity account of person identity. One comes to think now 'It is false that I will be in pain at *m*'. You change your mind about the truth of the thought 'I will be in pain at *m*' but "you have not thereby changed your understanding of the content of your own I-thought. Your conceptual grasp of what has to be the case for your I-thought to be true has not changed at all" (Nida-Rümelin 2010, 199). Only your view of what empirical criteria are necessary and sufficient for your I-thought to be true has changed. NR concludes that our conceptual grasp of what it is for one have a certain property at a future time is independent of any empirical criteria for transtemporal identity you accept.

> Claim 2. Transtemporal self-attribution is conceptually invariant with respect to changes in the thinker's accepted criteria of identity of people across time. (Nida-Rümelin 2010, 203)

This contrasts with individuals who are not conscious. Claims 1 and 2 (NR argues) entail claim 3:

> Claim 3: Transtemporal self-identification is conceptually invariant with respect to changes in a thinker's accepted criteria of identity of people across time. (Nida-Rümelin 2010, 203)

NR argues that given claims 1-3, it follows that the claims apply also for other-directed thought. If we think another is capable of first person thought, we will think that she will conceive of herself as a subject of properties in the future in a way that is conceptually fundamental relative to self-identification, and conceptually invariant with respect to views about empirical criteria for personal identity. In thinking about what it would be for another person to be one or the other of two people who result from a fission event involving her, we will think of it as a matter of what has to be so for her first person thought about her having future properties to be true. NR argues further that this extends to subjects of experience who are not capable of first person thoughts, such as infants and non-linguistic animals: we ask whether if such a subject *could* think first person thoughts, its first person thought about its future properties *would* be true, in asking about what individual it will be in the future. This then generates claims 4-6:

> Claim 4: Transtemporal attribution of properties to other experiencing subjects is conceptually invariant with respect to changes in the thinker's accepted criteria of subject identity across time.

> Claim 5: Transtemporal attribution of properties to others is conceptually prior to transtemporal identification with respect to others.

> Claim 6: The conceptual content of other-directed transtemporal identification is invariant with respect to possible changes of the accepted criteria of subject identity across time. Transtemporal criteria of subject identity do not enter the conceptual

content of other-directed transtemporal identification. (Nida-Rümelin 2010, 204-205)

The two most important claims here are that in thinking in the first person mode about what our future properties are (a) we have a *positive* understanding of what is involved (b) which is *independent* of our views with respect to change in the thinker's accepted criteria of subject identity across time. If this is true, then we cannot make use of the rejoinder to the fourth proposal considered above, because while it would capture an objective fact that would be something in addition to what is expressed by D, it would not be independent of empirical criteria for transtemporal identification.

The emphasis on a positive understanding in (a) is important. It is this I wish to examine. I grant in thinking of ourselves as having properties in the future in the "first person mode" we do *not* think of ourselves under a conception that presupposes empirical criteria of transtemporal identity. But the idea that we have a positive understanding of what is involved goes beyond this, and this, as we will see, is required if the argument is to go through.

What does the emphasis on positive understanding in (a) come to and why is it important? P1 and P2 differ only in that in P1 where the predicate '= LA' appears in P2 the predicate '=RA' appears. We might say that the positive understanding of the difference involved attaches to these, but (i) this would not give any special weight to the first person mode of thought and (ii) if this is all the positive difference comes to the proponent of empirical criteria of transtemporal identity could appeal to it as well. The positive understanding that carries the weight must then, I think, attach to the mode of presentation

of the self. Thus, the weight rests on the idea that we think in the first person mode of ourselves in a way that *positively* characterizes the self so as to *exclude* that the self has empirical criteria for transtemporal identity. It has to *exclude* that the self has empirical criteria of transtemporal identity, because if it leaves it open (if it were "topic neutral," for example), the argument for subject-body dualism collapses.[4]

If we add to this that

[MT] if we are identical with, composed in part, or constituted by any material object or stuff (or even immaterial stuff), then we have empirical criteria for transtemporal identity,

we can infer that we are not identical with, etc., any material object (or immaterial stuff). This would secure subject-body dualism. If we assume that necessarily subjects of experience are capable of the relevant sort of non-illusory first person thought, then we can conclude that subject-body dualism is necessarily true.

[4] A referee asked why NR would have to accept that the positive understanding excludes empirical criteria of transtemporal identity. Suppose it doesn't. Then whatever factual difference there is between P1 and P2 that rests on this positive understanding is compatible with empirical criteria of transtemporal identity, and therefore compatible with a materialist view, and therefore compatible with the rejection of subject-body dualism. So the argument for subject-body dualism won't go through if we leave it open that we have empirical criteria for transtemporal identity.

In light of this, we can see that the appeal to a *factual* difference between P1 and P2 is not essential for the argument. It rests on an assumption that will drive the argument independently. We can state the argument as follows:

1. We think (correctly) in the first person mode of ourselves in a way that positively characterizes the self so as to exclude that the self has empirical criteria for transtemporal identity.
2. If we are identical with, composed in part, or constituted by any material object or stuff (or even immaterial stuff), then we have empirical criteria for transtemporal identity. [MT]
3. We are not identical with any material object [1-2].

Call this the fundamental argument.

3. First person thought

I agree that the fundamental argument is valid, but reject the first premise. The first premise rests on claim 2. I reject claim 2, repeated here, on the reading that supports the first premise.

> Claim 2. Transtemporal self-attribution is *conceptually* invariant with respect to changes in the thinker's accepted criteria of identity of people across time. (Nida-Rümelin 2010, 203)

I accept a weaker claim, however. The weaker claim is 2*.

Claim 2* The content of transtemporal self-attributions is invariant with respect to changes in the thinker's accepted criteria of identity of people across time.

This is distinct from Claim 2, which I take to entail, via the modifier 'conceptually', that there is a positive way in which we present the self that excludes criteria of identity of people over time. (If we say that claim 2 and 2* are the same, then the objection is that premise 1 doesn't follow.)

How can claim 2* be true though claim 2 is not? (Alternatively, how can claim 2* be true yet premise 1 not follow?) The answer is that we could think of ourselves directly, without bringing ourselves under any concept, or conception, at all, other than, perhaps, that of a thing, and think of ourselves directly as being related to a time in the future of the present by having a certain property then. This would explain why what we are thinking is invariant with respect to our criteria for transtemporal identity over time. To put it in other words, we do not pick out the future self as

the *x* such that x exists at future time *t* and *x* bears *R* to *me* at the present time,

and attribute a property to whatever is denoted by that, but we think:

at some time *t* in the future of the present, *I* [thought of directly at the present time] have such and such a property at *t*.[5]

To take an example, think about the proposition expressed by the English sentence 'I will have a headache tomorrow morning', as asserted by me now, taking 'I' to introduce into the proposition expressed just its referent, i.e., the speaker. The proposition is a singular proposition. The rule for the use of the pronoun involves a description:

(R) For any *x*, any time *t*, any subject *u*, 'I' refers to *x* at *t* as uttered by *u* iff *x* = *u*.

But the rule doesn't enter into the content of the proposition, only what object it assigns as the referent of 'I'. Take '@' to be a directly referring term that picks out me.[6] Then, where '*N*' directly refers to the time of utterance (and '>' means 'is later than'), the proposition is:

[5] I made this claim in note 3 of the 2013 paper, and related it to the final point I made in that paper. NR's reply has helped me to see that this is where the most fundamental disagreement between us lies.

[6] The reader will notice that I am not distinguishing *de re* and *de se* thoughts. I give an explanation for the substitution puzzles that motivate drawing the distinction in (Ludwig 1996). My view is that when one uses the first person pronoun, given knowledge of the rule, one uses it on the basis of thoughts that are indeed directly about the self, but when we use a proper name, even if it has the same referent, there is no guarantee that one locates the referent directly in thought. This is what gives rise to the substitution puzzles. The crucial point for the argument in the text is just that we do think of ourselves directly (in thought) when we express a thought using the first person pronoun, this is an attitude toward a singular proposition, and this expresses what we have in mind by the first person mode of thought.

[FH] [$\exists t$: $t > N$](@ has a headache at t and t lies within the morning of the tomorrow of N)

That is clearly invariant with respect to @'s views about transtemporal criteria for personal identity. If the content of the thought I have when I express myself using the sentence 'I will have a headache tomorrow morning' is given by the proposition that this sentence expresses, then while claim 2* is true, claim 2 (interpreted as sketched above) is not (because as interpreted above, claim 2 says we present ourselves in first person thought in some positive manner, as having some features, whereas if the thought about the self is direct, the self is not presented as having any features). Likewise then premise 1, repeated here, of the fundamental argument is false. What is true is premise 1′.

1. We think (correctly) in the first person mode of ourselves in a way that positively characterizes the self so as to exclude that the self has empirical criteria for transtemporal identity.

1′. We think (correctly) in the first person mode of ourselves in a way that does not include that the self has empirical criteria for transtemporal identity.

1′ is compatible with our having empirical criteria for transtemporal identity.

What are the implications for claims 1 and 3, repeated here?

> Claim: 1: Transtemporal self-attributions are conceptually prior to self-identifications. (Nida-Rümelin 2010, 203)

> Claim 3: Transtemporal self-identification is conceptually invariant with respect to changes in a thinker's accepted criteria of identify of people across time. (Nida-Rümelin 2010, 203)

In thinking about oneself as having a property in the future, as in [FH], one thinks of oneself directly at the present time, and of oneself so thought of as related by having a headache then to a time located in the future of the present time. This guarantees that the person one is thinking of as having a headache at a time subsequent to the present is oneself. But does this show that self-attributions are conceptually prior to self-identifications?

What does this mean? The natural reading is that one can self-attribute future properties without first identifying some future self as oneself. If this is what it means, then we can accept it, but it does not get us very far, if the present picture is correct. For on that picture, while it is true that when one thinks of oneself as having a property in the future, there is no question that arises about identifying oneself as the one that one is thinking about, this is just because one is thinking about oneself directly in the present as being related to a future time which one picks out by a restricted quantifier anchored by a direct reference to the time of the thought. One has not *identified* oneself as some future individual. One has only picked oneself out in the way one does when thinking a thought about a property one has at the present moment. One is not called on to think in any substantive way about what it would be for one to survive to have a property in the future. And self-attributions of future properties in this way leaves it open what would have to be true for anything at any future time to be oneself.

What about claim 3, though? This has to be given up if claim 2 is given up because it would, like claim 2, presuppose that in thinking about our future selves in the first person way we have a positive conception of the self which rules out our having empirical criteria of transtemporal identity.

These remarks carry over to claims 4-6.

So far, I have only said what follows if we accept this sketch of the content of first person thoughts about future properties of the self. One might object at this point to the account of the content of the proposition expressed by 'I will have a headache tomorrow' or to the claim that the thought I have about properties of my future self expressed with the first person pronoun in subject position is properly specified by the proposition expressed by the sentence that I use. I respond to this in two stages. First, I give an argument to show that we do not pick ourselves out via any purely qualitative mode of presentation, so that there must be some element of direct reference in thought to the self. Second, I argue that, given this, we have no reason to think we subsume the self under any positive conception or concept in referring to ourselves in the "first person mode."

The argument for the claim that we do not pick ourselves out via any purely qualitative mode of presentation, goes as follows (Ludwig 1996).

1. We know that we are able to think about ourselves and attribute properties to ourselves.
2. If we were able to think of ourselves only by way of a purely qualitative mode of presentation (or description), then we would not know that we are able to think about ourselves and attribute properties to ourselves.

3. Therefore, we do not pick ourselves out only by way of a purely qualitative mode of presentation (or description).

The subargument for premise 2 goes as follows.

1. We do not know that the universe does not contain (timelessly speaking) qualitative duplicates of everything that exists (or at least of ourselves up to the extent of our knowledge).
2. If the universe contains qualitative duplicates of everything that exists (or at least ourselves up to the extent of our knowledge), then no purely qualitative mode of presentation (or description) uniquely denotes any individual (or ourselves to the extent of our knowledge).
3. Therefore, if we were able to think of ourselves only by way of a purely qualitative mode of presentation (or description), we would not know that we are able to think about ourselves and attribute properties to ourselves.

The argument for premise 1 of the subargument is that we do not know that, for example, Nietzsche's hypothesis of eternal recurrence is not true, that is, we do not know that the universe does not repeat each temporal segment of it qualitatively identically an infinite number of times.[7]

[7] I say 'for example', because this is not the only hypothesis we could appeal to here. Do we know that there are not an infinite number of spatio-temporally isolated universes (like David Lewis "possible worlds" but without the commitment to explaining modal claims in terms of them) among which there are qualitative

It might be said in response that physics tells us that Nietzsche's hypothesis is in fact false.[8] The second law of thermodynamics ensures that the universe will die a heat death in a state of maximum entropy. However, this overstates what we know. The matter is not entirely settled in physics whether the universe iterates through infinite cycles (a Big Bang followed by a Big Crunch, followed by a Big Bang and so on).[9] But even granting that it is settled that cycles were not physically possible—that the universe is open and will die a heat death—this doesn't matter. The fact is that we do not need to know what physicists know (if they know that) in order to know that we are able to refer to ourselves (see note 7 also). For most people, it is epistemically open that there are qualitative duplicates of them, but this is not a threat to their knowledge that they can refer to themselves. So the real force of the argument is that our knowledge that we think about ourselves is not hostage to whether there are qualitative duplicates of us, but this could be so only if we

---

duplicates of ours? We could not rule these out by any empirical means, and since it is possible that a universe contain an infinite number of spatio-temporally isolated universes one of which is just like ours, there is no a priori argument to show that this hypothesis is false.

[8] Eternal recurrence does not follow form the hypothesis of infinite time. Consider two cylinders with marks that line up at time zero, one of which begins rotating once per second and the other $\sqrt{2}$ times per second. They will never line up again in the original configuration, since that would require $\sqrt{2}$ to be equal to the ratio of two integers. This is a variant of a counterexample provided by Georg Simmel in 1907 and reported in (Kaufmann 1974, 327). What is at issue is not the necessity but the epistemic possibility of eternal recurrence.

[9] See (Penrose 2010).

could think about ourselves directly and not only as the unique possessors of some set of (purely qualitative) properties.

This leaves it open that we think of ourselves by a mode of presentation that functions like a complex demonstrative. Let 'subject' express the special conception of the type of being that represents a positive conception of the self that excludes empirical criteria of transtemporal identity. Then it is open still that we think about ourselves in a thought of the form 'that subject is in pain'. We can think of 'that subject' as a restricted quantifier of the form 'the *x*: *x* is *that* and *x* is a subject' as suggested in (Lepore 2000), or as we could think of the function of 'subject' as a filter on how the demonstrative element refers, as Kaplan suggested (Kaplan 1989, 515). In either case, we refer directly to the self, but at the same time in doing so bring the self under the concept of a subject. The second of these components, though, does not play a crucial role in the mechanism of reference itself.

My argument against first person thought involving this extra component has three parts. First, (a) it is gratuitous and (b) we should not adopt views that are gratuitous. It is gratuitous because the main diagnostic for there being a positive conception of the self is that in self-attributing future thoughts we think of the self in a way that is invariant with respect to changes in our conception of empirical criteria for transtemporal identity. But once we see that we refer to the self directly, we have an explanation for this that does not require that we bring the self under any positive conception. Second, we could respond to the charge that it is gratuitous if there were present in reflection on self-attributions of properties some positive conception of the self that did exclude empirical criteria of transtemporal identity. But we do not, in fact, bring ourselves under such a positive conception of the self in self-attributing properties. Or, to speak more cautiously, I do not

find, that when I rap my knuckles on my desk, and have the thought that I am in pain, I am in thinking *I qua subject am in pain*, where the concept of a subject is a positive conception of what I am thinking about. It is implied by my thinking about myself at all that I am a thinking thing, but this does not enter into the content of the thought itself, apart from its "predicate". Even if it did, it would not be the right sort of conception to support the exclusion of empirical criteria for transtemporal identity. Third, if we did bring the self under a positive concept that excluded empirical criteria of transtemporal identity, then we should be able to articulate what about it precludes our having them. What is it though? Is it that we are composed of immaterial stuff rather than material stuff? This won't work because the same argument applies here, as NR notes. If we were some immaterial stuff, there would be criteria for transtemporal identity appropriate for it, but it is clear that the content of self-attributions of future properties would be invariant with respect to our conceptions of what that involved. Only thinking of the self as a thing that is utterly simple, and whose persistence through time is not governed by informative criteria at all, would seem to have the right character. This would not rule out (a priori) its being a material thing that was simple, however. And it is hardly clear that when we think about ourselves we are thinking about something that is utterly simple.[10] This seems to be an *open* question relative to our thinking of ourselves in the first person mode, in the way we express when we use the first person pronoun.

[10] This point is connected with the remark in note 10 in (Ludwig 2013) that the original argument applies only to material objects that are subject to fissioning. Absolutely simple objects are not capable of fissioning.

Finally, even if we did invariably bring the self under such a concept, given that it is not required for us to refer to ourselves in thought in the first person mode, it would remain an open question whether we were correct to bring ourselves under such a concept, and this would not require that we think in general that we mistakenly attribute thoughts to ourselves, for the illusion would extend only to what is from the point of view of our ordinary attributions an extraneous and unnecessary addendum to first person thoughts.

To summarize: in first person thought we think about ourselves directly at the present time. This is true also when we attribute to ourselves properties in the future, because we are thinking of ourselves as picked out now directly as being related to times in the future (of the time of the thought) by way of having various properties then. Thus, the content of those thoughts are invariant with respect to variations in our conceptions of criteria for transtemporal identity of people. This does not require that we have a positive concept of the self that excludes empirical criteria for transtemporal identity. The suggestion that we do bring the self under such a concept is not supported by reflection on first person thought, and in the absence of that it is gratuitous to suggest that we bring the self under such a concept. Finally, even if we did, it would not show that we did not have empirical criteria for transtemporal identity and whatever illusion this involved would be localized and not undermine the vast majority of what we think about ourselves. The fundamental argument fails.

4. Possibilities and the Third Interpretation

I turn now to the question whether the third interpretation of the factual difference between P1 and P2 supports subject-body dualism. For convenience, I repeat P1-P3 here.

P1: D and A = LA.

P2: D and A = RA.

P3: D and A ≠ LA and A ≠ RA.

The third interpretation was that P1 and P2 are both genuine possibilities. On this interpretation, the second premise of the streamlined master argument can be reformulated as in 2′ .

2′. P1 and P2 could be genuine possibilities only if subject-body dualism were true.

In (Ludwig 2013), I argued (in part) that: for the possibility that P1 or the possibility that P2 to show that as a matter of fact A did not have empirical transtemporal identity conditions, it would have to be necessary that in every possible world either A = LA or A = RA; but this is incompatible with the assumption that it is possible that A ≠ RA *and* A ≠ LA, and if that last were true at the actual world, it would be compatible with our having empirical criteria of transtemporal identity in the actual world.

NR responds by introducing the assumption that we have our transtemporal criteria for identity essentially. If there is a possible world in which A = LA (when RA is also produced from A by a fission event), in that world A does not have empirical transtemporal identity conditions. If persons have their transtemporal identity conditions essentially, it follows that in the actual world A does not have empirical transtemporal identity conditions.

Even granting the assumption, there is a problem with the third interpretation. I introduced this problem in the (2013) paper as a response to the first objection. That relied on the charge that P3 was supposed to be possible. But if it is, then, given the necessity of identity, P3 rules out P1 and P2. I replied that the rejoinder was too powerful, because it also means that P1 and P2 cannot *both* be possibilities. It would have been more straightforward, perhaps, to say that the basic problem with the third interpretation is simply that it is incoherent to maintain that P1, P2, and P3 are all possibilities. Suppose that P1 is possible. Then, given the necessity of identity, (NI),

> (NI) For any x, y, if $x = y$, then for any possible world w, if x exists in w or y exists in w, then $x = y$ in w

if it is possible that A = RA, then in any world in which A or RA exist, A = RA, and since RA ≠ LA, it follows that it is not possible that A = LA. And if it is possible that A = LA, then it follows that it is not possible that A = RA. And if it is possible that A ≠ RA *and* A ≠ LA, then it follows that it is not possible that A = RA and it is not possible that A = LA. We cannot hold that each of P1, P2 and P3 are genuine possibilities. Only one of them can be. Thus, the third interpretation of the factual difference in fact involves a claim that is false, namely, that P1 and P2 are both genuine possibilities.[11]

[11] It might be suggested that NR needs only one of the two possibilities, not both. So she might say: P1 is possible but not P2 or P2 is possible but not P1. But this is not the argument that NR is advancing, and, it seems, for good reason. For the problem with this that whatever might ground the claim that P1/P2 is possible would seem to ground the claim that the other is possible, given the symmetry of the setup. In

P1, P2 and P3 strike us as prima facie possibilities, but they cannot all be genuine conceptual or metaphysical possibilities. In what sense are they possibilities? The answer, I think, and this will now connect the discussion of this interpretation with the discussion in the previous section, is that P1, P2 and P3 are epistemic, not metaphysical or conceptual possibilities. They are epistemic possibilities because when we think of ourselves in the first person mode we do so directly, without thereby revealing what our natures are.

We can imagine what it would be like to be one or the other of the two successors in a fission scenario by self-attributing future experiences appropriate for being the one or the other. We might, for example, be told that the right successor will prick the finger of her right hand while the left will prick the finger of her left hand. We can think of what it would be like to be in the position of the one, feeling the right finger pricked, rather than the left, or vice versa. In thinking of this, in the first person mode, imagining that one is feeling the prick tomorrow *there*, one thinks a thought that does not involve any incoherence. It is presented as an epistemic possibility at least relative to the content of the thought. This is not directly to think of ourselves as PR or PL. But it is a short step. For what seems compatible with D is that we are having experiences of a sort which only PR or which only PL would be having. Thus, it will seem open (relative to the content of the thought) that we are PR or PL, respectively. And when we think of others faced with fissioning, we can imaginatively project ourselves into their shoes and see that they can conceive the same thing from their point of view. It is epistemically open for them as well, relative to the content of their future directed thoughts, that they are the one or the other.

---

addition, it amounts to the rejection of P3, which is the position of the materialist, and so without further argument it would be question begging.

However, (a) these cannot be more than epistemic possibilities because they cannot both be metaphysical or conceptual possibilities given (NI), and (b) we have an explanation of why they are epistemic possibilities for us that does not require that they be genuine metaphysical or conceptual possibilities. For in thinking or imagining ourselves as feeling such and such a prick *there*, we have a thought of the form ϕ(I), where I use 'I' to represent the unmediated thinking of the subject of the thought by its subject. The nature of the object of thought is not presented in the thought. Consequently it appears open to us that it be true when all the conditions specified by the fissioning scenario are in place. But what we are picking out may be a material object even if not so presented. Consequently what we are thinking, coherently so far as the content of the thought goes, may not be a genuinely possibility, because as a matter of fact we have empirical conditions for transtemporal identity.

Our epistemic position with respect to our natures in first person thought is analogous to our position with respect to the natures of stuffs for which we introduce natural kind terms. We attach natural kind terms to natural kinds through application to examples we pick out by how they present themselves to us. But the kind properties that we aim to keep track of are not given by the features by which we pick them out. So the thoughts we entertain about them (prior to discovering what the kind property is) do not reveal their natures to us. So too in thinking of ourselves in the first person mode, how we pick out ourselves does not reveal what we are, and so it is epistemically open what sort of thing we are. The difference is that we are not picking ourselves out by any features we have, so that the self is not presented to us in the first person mode of thought in any positive way whatsoever. All of that lies in what we predicate of the self.

To summarize, NR is right that requiring that we have our transtemporal identity conditions essentially together with the assumption that, for instance, P1 is possible, entails that we are not material things or constituted from material things. But even granting the premise, this doesn't rescue the argument on the third interpretation of *factual difference*. For the third interpretation asserts that P1 and P2 (and in fact P3) are all genuine possibilities. But this is impossible given (NI). P1, P2 and P3 are rather epistemic possibilities. That they are epistemic possibilities is explained by the fact that in first person mode future attributions of properties we pick out ourselves directly, and so in a way that does not present itself as in conflict with our being one or the other of the successors in a fission case because it is silent on what our natures are and so on whether we have empirical criteria of transtemporal identity.

5. Summary and Conclusion

I have argued that the fourth interpretation that NR offers of "factual difference" in her (2010) argument for subject-body dualism does not secure an interpretation on which the argument goes through, but that a more fundamental claim motivates this way of putting the factual difference that carries the argument by itself, namely, that first person thought involves a positive conception of the self that excludes our having empirical criteria of cross-temporal identity. In response, I argued that in first person thought, we think of ourselves directly. We do not present ourselves under a special concept that rules out our having transtemporal identity conditions. Thus, the crucial premise in the underlying argument is mistaken. I argued further that the third interpretation, that P1 and P2 are both genuine possibilities, is incompatible with the necessity of identity. They are instead

epistemic possibilities that are explained by our primary reference to ourselves being direct in a way that does not reveal what sort of thing we are.

References

Kaplan, David. 1989. Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives. In *Themes From Kaplan*. New York: Oxford University Press.

Kaufmann, Walter. 1974. *Nietzsche: Philosopher, Psychologist, Antichrist*. 4th ed. Princeton: Princeton University Press.

Lepore, Ernest., & Ludwig, Kirk. (2000). The Semantics and Pragmatics of Complex Demonstratives. *Mind*, 109(434), 199-240.

Ludwig, Kirk. 1996. Singular Thought and the Cartesian Theory of Mind. *Noûs*, 30(4), 434-460.

Ludwig, Kirk. 2013. The Argument for Subject-Body Dualism from Transtemporal Identity. *Philosophy and Phenomenological Research, 86*(3), 684-701.

Nida-Rümelin, Martine. 2006. *Der Blick von Innen: Zur Transtemporalen Identität Bewusstseinsfähiger Wesen*: Suhrkamp.

———. 2010. An Argument from Transtemporal Identity for Subject-Body Dualism. In *The Waning of Materialism*, edited by G. Bealer and R. Koons. Oxford: Oxford University Press.

———. 2013. The Argument for Subject Body Dualism from Transtemporal Identity Defended. *Philosophy and Phenomenological Research* 86 (3):702-714.

Penrose, Roger. 2010. *Cycles of Time: an Extraordinary New View of the Universe*. London: Bodley Head.