Copula

Vol. XXVII June 2010

Journal of the Department of Philosophy

Department of Philosophy
Jahangirnagar University
Savar Dhaka Bangladesh

Copula, Vol. XXVII, June 2010, Jahangirnagar University

The Impossibility of Mind-swapping sort of Thought Experiments in the Study of Personal Identity

Mostofa Nazmul Mansur

[Abstract: Bernard Williams in his paper "The Self and the Future" argues that mind-swapping sort of thought experiments should not be considered as true guides in determining the criterion of personal identity, because one can construct this sort of thought experiments that supports the 'mentalistic' criterion thesis of personal identity and at the same time one can construct a similar mind-swapping sort of thought experiments that contains the same theoretical strength which supports the 'bodily continuity' criterion thesis of personal identity. Inspired by the main spirit of Williams' project, in the present paper I have tried to show that thought experiments based on memory exchange are practically and theoretically impossible, and hence, any account of personal identity which involves such thought experiments are fundamentally mistaken.]

Thought experiments have been used in philosophy and sciences for centuries. In the study of personal identity (PI) thought experiments seem to be the most used tool in finding the criterion of PI. Bernard Williams, in his paper "The Self and the Future," shows two principle limitations of using thought experiments in the study of PI. My paper focuses on the second limitation that Williams mentioned. To make this limitation visible Williams offers a set of mind-swapping or body-swapping sort of thought experiments which provides support for 'mentalistic' criteria (i.e. thoughts, memory, character traits) thesis of PI. Then he offers another set of thought experiments which, contrary to the first set of thought experiments, provides support in favor of 'bodily continuity' criterion thesis of PI. So, we see, these sorts of thought experiments, in the study of PI, lead us to two contrary conclusions. Thus the lesson from these thought experiments is that mind-swapping or body-swapping sort of thought

experiments should not be considered as true guides in determining the criterion of PI. I agree with this lesson. To support this position, in this paper I will add some of my observations regarding the issue. My key observation is that mind-swapping or body-swapping sort of thought experiments based on memory exchange are neither practically nor theoretically possible. So, any account of PI which involves such thought experiments are fundamentally mistaken.

First, let us consider Williams' first set of thought experiments. (pp. 181-185). Here Williams asks us to imagine two persons, *A* and *B*, who are about to undergo an unusual experiment. Before the experiment takes place, both *A* and *B* have been informed that all the thoughts, memories and character traits of *A* will be extracted and, then, will be transmitted to *B*; and the same thing will happen to *B*. That is, after the experiment we will have:

(i) an *A-body-person* with *B*'s thoughts, memories and character traits
and, (ii) a *B-body-person* with *A*'s thoughts, memories and character traits.

Now, suppose, before the experiment starts *A* and *B* are told that one of the two resultant persons will be awarded $ 100,000 whereas the other will be tortured. Both *A* and *B* are asked to decide on purely self-interested grounds whether the reward should go to the *A-body-person* or to the *B-body-person*. Suppose again, pre-experimental *A* prefers the *A-body-person* to be tortured and the *B-body-person* to be rewarded after the experiment. And, pre-experimental *B* chooses conversely. Since the experimenter cannot act in accordance with both of these sets of preferences, he must act in accordance of either pre-experimental *A*'s set of preferences or pre-experimental *B*'s set of preferences. Suppose the experimenter decides to act in accordance of pre-experimental *A*'s set of preferences, i.e. after the experiment he rewards the *B-body-person* and tortures the *A-body-person*. Now, when the experiment is over and promised reward and torture are distributed, the *B-body-person*, whose thoughts, memories and character traits correspond the pre-experimental *A*, will rightly say, "That's just the outcome I chose to happen"; and the *A-body-person*, whose thoughts, memories and character traits correspond the pre-experimental *B*, will say, "That's not the outcome I chose to happen." (p. 182).[1] These reactions prove that after the experiment:

(i) the pre-experimental *A* has become *B-body-person*
and, (ii) the pre-experimental *B* has become *A-body-person*.

That means that a mind-swapping or a body-swapping has been occurred, and by the exchange of mentalistic events, such as thoughts, memories and character traits, an exchange of *personhood* has been happened. This feature provides strong support for the 'mentalistic' thesis of PI because the subjects of the experiment, e.g. *A* and *B*, identify themselves with

their mentalistic events, not with their bodily continuity. This argument corresponds a valid form of Modus Ponens:

P_1: If one identifies oneself with her thoughts-memories-and-character traits despite 'body-swapping,' then thoughts-memories-and-character traits are the criteria of PI.

P_2: [There are thought experiments which show that] One identifies oneself with her thoughts-memories-and-character traits despite 'body-swapping.'

C: Therefore, thoughts-memories-and-character traits are the criteria of PI. (That is, 'mentalistic' thesis of PI is true.)

However, Williams then offers three other thought experiments of the same sort with slight changes in *A* and *B*'s preferences, each of which reconfirms the conclusion of the original argument. Thus, Williams says:

> [A]ll the results suggest that the only rational thing to do, confronted with such an experiment, would be to identify oneself with one's memories, and so forth, and not with one's body. The philosophical arguments designed to show that bodily continuity was at least a necessary condition of personal identity would seem to be just mistaken. (p. 185).

Now consider Williams second set of thought experiments. (pp. 186-89). Here Williams actually offers one thought experiment, and then, restates the same though experiment dividing it into a series of six interconnected cases. However, the second scenario runs this way: a person *A* is told that he is going to be tortured tomorrow. Fear is the natural reaction to such an announcement. Then the person is told again that there is nothing to fear because when the time of torture comes, he will not remember that he was told about the torture beforehand since he will undergo a powerful amnesia shortly before the torture takes place. But, Williams rightly claims, this assurance does not comfort the person because the person will consider it a torture upon him. It does not matter whether he forgot it or not. Then, again, the person is told that not only will he not remember that he is to be tortured, but he will not remember many things he is now in a position to remember. This will not relax the person *A* because one should not like to find oneself in a completely amnesic state. Now the person is told again that not only will he not remember the things that he remembers now, he will, rather, be given a completely different set of thoughts, memories and character traits. Williams suggests that instead of reducing the fear this announcement will increase the horror because the person now may begin to think that he will be tortured and at the same time he will become such a mad person who will think of him as George IV or somebody else other than himself. (p. 186). But what will happen if the person is told that his new set of thoughts, memories and character traits will be identical with those of another person

their mentalistic events, not with their bodily continuity. This argument corresponds a valid form of Modus Ponens:

P_1: If one identifies oneself with her thoughts-memories-and-character traits despite 'body-swapping,' then thoughts-memories-and-character traits are the criteria of PI.

P_2: [There are thought experiments which show that] One identifies oneself with her thoughts-memories-and-character traits despite 'body-swapping.'

C: Therefore, thoughts-memories-and-character traits are the criteria of PI. (That is, 'mentalistic' thesis of PI is true.)

However, Williams then offers three other thought experiments of the same sort with slight changes in *A* and *B*'s preferences, each of which reconfirms the conclusion of the original argument. Thus, Williams says:

> [A]ll the results suggest that the only rational thing to do, confronted with such an experiment, would be to identify oneself with one's memories, and so forth, and not with one's body. The philosophical arguments designed to show that bodily continuity was at least a necessary condition of personal identity would seem to be just mistaken. (p. 185).

Now consider Williams second set of thought experiments. (pp. 186-89). Here Williams actually offers one thought experiment, and then, restates the same though experiment dividing it into a series of six interconnected cases. However, the second scenario runs this way: a person *A* is told that he is going to be tortured tomorrow. Fear is the natural reaction to such an announcement. Then the person is told again that there is nothing to fear because when the time of torture comes, he will not remember that he was told about the torture beforehand since he will undergo a powerful amnesia shortly before the torture takes place. But, Williams rightly claims, this assurance does not comfort the person because the person will consider it a torture upon him. It does not matter whether he forgot it or not. Then, again, the person is told that not only will he not remember that he is to be tortured, but he will not remember many things he is now in a position to remember. This will not relax the person *A* because one should not like to find oneself in a completely amnesic state. Now the person is told again that not only will he not remember the things that he remembers now, he will, rather, be given a completely different set of thoughts, memories and character traits. Williams suggests that instead of reducing the fear this announcement will increase the horror because the person now may begin to think that he will be tortured and at the same time he will become such a mad person who will think of him as George IV or somebody else other than himself. (p. 186). But what will happen if the person is told that his new set of thoughts, memories and character traits will be identical with those of another person

who is now alive? Williams claims that it will not give the person any reason not to fear the torture. It will still be rational to fear the upcoming torture. So, here we find:

(i) a total change of the person *A*'s mentalistic states has been occurred and (ii) the person *A*'s fear of future bodily torture still continues.

Now, if these assessments of the second thought experiment are correct, then they provide strong support for the 'bodily continuity' thesis of PI because here, we see, the person *A* identifies himself with his bodily continuation despite total changes in his mentalistic states. Like the first one, this argument also corresponds to a valid form of Modus Ponens:

P_1: If one's fear of bodily torture continues despite the total change of psychological phenomena, then bodily continuity is the criterion of PI.

P_2: [There are thought experiments which show that] One's fear of bodily torture continues despite the total change of psychological phenomena.

C: Therefore, bodily continuity is the criterion of PI.
(That is, 'bodily continuity' thesis is true.)

Thus, here we reach to a conclusion which is contrary to the conclusion of the first argument (first set of thought experiments). So, Williams is careful about this second thought experiment and re-evaluates it listing six possible things that the person *A* could be told to see whether there was any step where *A* could rationally get rid of the fear of torture. He sets six possibilities as follows (pp. 190-96):

(i) A is subject to an operation which produces total amnesia;

Williams reasonably claims that everyone should agree that *A* will fear the torture in this situation.

(ii) amnesia is produced in A, and other interference leads to change in his character;

This situation does not provide any good reason to feel reassured about *A*'s torture the next day.

(iii) changes in his character are produced, and at the same time certain illusory 'memory' beliefs are induced in him; these are quite fictitious kind and do not fit the life of any actual person;

Instead of reducing the fear, according to Williams, this situation may increase the fear because the person *A* may think that he is going to be tortured the next day and at the same time he is going to be a mad person who will identify himself with some fictitious person.

> *(iv) the same as (iii), except that both the character traits and the 'memory' impressions are designed to be appropriate to another actual person;*

This situation is significant. It introduces a new person *B*. However, according to Williams, this introduction of another actual person, *B*, does not make any material difference—if one fears of torture *upon oneself* in the situation (iii) despite having new 'memory' impressions, then there is no reason not to fear if these new 'memory' impressions have a model in the actual world. (p. 191).

> *(v) the same as (iv), except that the result is produced by putting the information into A from the brain of B, by a method which leaves B as he was before;*

According to Williams there is no reason not to fear in situation (v) either because if someone fears torture in situation (iv) despite having new 'memory' impressions which have a model in the actual world, then there is no reason for him not to fear if these new 'memory' impressions are copied from that model in question. (p. 191).

> *(vi) the same happens to A in (v), but B is not left the same, since a similar operation is conducted in the reverse direction;*

This is the situation where the upholders of the 'mentalistic' thesis of PI may argue that the person *A* gets rid of the alleged fear here because he should identify himself in this situation as the person *B* in the person *A*'s body. But Williams does not agree. He claims that there is no significant difference between (v) and (vi). What is added in situation (vi) is totally external, and in that sense irrelevant, to *A* of (v). That is, if *A* is *A-body-person* in (v) and reasonably fears the torture, then *A* must be *A-body-person* in (vi) and should fear the torture because what is happening to *B* does not matter to *A* or *A*'s *personhood*. (pp. 192, 196). Again, if *A-body-person* is not really *A* in (vi), then *A-body-person* is not *A* in situation (v) either because they are exactly the same. And this way, *A-body-person* is not really *A* in situations (iv), (iii), (ii) and (i) because there is no significant difference between every two consecutive situations. But it is not possible—nobody will agree that, at least, in (i) and (ii) the *A-body-person* is not really *A*. If such is the case, then *A-body-person* is really *A* in (i) and (ii) and has reason to fear torture the next day, and consequently, the same is true of all other situations. (p. 192).

From the above assessments of the 'six possibilities' we see that it is rational for *A* to think that *A-body-person* will be *A* despite total change of mentalistic events, such as thoughts, memories and character traits. That means that we recognize our identity as something fundamentally bodily and that is why our fear of bodily torture extends through any amount of psychological disruption. Thus, again, it is shown that the second

set of thought experiments provides support for 'bodily continuity' thesis of PI.

So, we see, Williams first offers a set of thought experiments that supports 'mentalistic' thesis of PI, and then, he offers another similar[2] set of thought experiments that supports 'bodily continuity' thesis of PI. That is, such thought experiments provide us different conclusions which are mutually exclusive. The lesson we learn from Williams' demonstrations of these two sorts of thought experiments is that mind-swapping or body-swapping sort of thought experiments based on memory exchange are fatally flawed ways to reaching answers about PI. So, these sorts of thought experiments should not be considered as guides in determining the criteria of PI.

Of course, opponents of Williams' project might argue that Williams' thought experiments are designed in a *subject-dependent* way (by subject I mean the subjects of the experiments, such as the persons *A* and *B*). They might say that the alleged fear, in the second set of thought experiments, extends from situation to situation despite total change of mentalistic events because the person *A* has a general intuition that bodily continuity is identical with the continuation of personhood. Otherwise, *A* would get rid of the fear of physical torture at situation (iii) where he was assured that he would get a new set of mentalistic events instead of the previous set of those. If any fear continued at and after that situation, that fear would not be the fear of physical torture but the fear of 'identity crisis' from the 'mentalistic' point of view, because situations (iii), (iv), (v) did not assure him that *he* would be the *only person* who would have that new set of mentalistic events. Situation (vi) assured that *he* would be the only person having that certain set of mentalistic events. And, in this situation the person, unless he had firm belief in the 'bodily continuity' thesis of PI, would completely get rid of any sort of fear of torture. Thus, what would be the outcome of Williams' second set of thought experiments depends on what general intuition the subject of experiment previously had. The same, the opponents of Williams' project might say, is true of Williams' first set of thought experiments. The person *A* (post-experimental *B-body-person*) and *B* (post-experimental *A-body-person*) had strong beliefs in the truth of 'mentalistic' thesis of PI. Otherwise, they would remember that they had gone through a body-swap and would acknowledge that they did not have their *own bodies*. That is, post-experimental *A-body-person* would say, 'it is my memory, but *not my body*.' And post-experimental *B-body-person* would say something similar. And this 'not my body' claim would severely attenuate the strength of the outcome of Williams' first set of thought experiments. So, the opponents might claim that like his first set of thought experiments, Williams' second set of thought experiments is subject-dependent.

set of thought experiments provides support for 'bodily continuity' thesis of PI.

So, we see, Williams first offers a set of thought experiments that supports 'mentalistic' thesis of PI, and then, he offers another similar[2] set of thought experiments that supports 'bodily continuity' thesis of PI. That is, such thought experiments provide us different conclusions which are mutually exclusive. The lesson we learn from Williams' demonstrations of these two sorts of thought experiments is that mind-swapping or body-swapping sort of thought experiments based on memory exchange are fatally flawed ways to reaching answers about PI. So, these sorts of thought experiments should not be considered as guides in determining the criteria of PI.

Of course, opponents of Williams' project might argue that Williams' thought experiments are designed in a *subject-dependent* way (by subject I mean the subjects of the experiments, such as the persons *A* and *B*). They might say that the alleged fear, in the second set of thought experiments, extends from situation to situation despite total change of mentalistic events because the person *A* has a general intuition that bodily continuity is identical with the continuation of personhood. Otherwise, *A* would get rid of the fear of physical torture at situation (iii) where he was assured that he would get a new set of mentalistic events instead of the previous set of those. If any fear continued at and after that situation, that fear would not be the fear of physical torture but the fear of 'identity crisis' from the 'mentalistic' point of view, because situations (iii), (iv), (v) did not assure him that *he* would be the *only person* who would have that new set of mentalistic events. Situation (vi) assured that *he* would be the only person having that certain set of mentalistic events. And, in this situation the person, unless he had firm belief in the 'bodily continuity' thesis of PI, would completely get rid of any sort of fear of torture. Thus, what would be the outcome of Williams' second set of thought experiments depends on what general intuition the subject of experiment previously had. The same, the opponents of Williams' project might say, is true of Williams' first set of thought experiments. The person *A* (post-experimental *B-body-person*) and *B* (post-experimental *A-body-person*) had strong beliefs in the truth of 'mentalistic' thesis of PI. Otherwise, they would remember that they had gone through a body-swap and would acknowledge that they did not have their *own bodies*. That is, post-experimental *A-body-person* would say, 'it is my memory, but *not my body*.' And post-experimental *B-body-person* would say something similar. And this 'not my body' claim would severely attenuate the strength of the outcome of Williams' first set of thought experiments. So, the opponents might claim that like his first set of thought experiments, Williams' second set of thought experiments is subject-dependent.

I, however, do not agree that this allegation of subject-dependency discredits Williams' project. We must consider that the main aim of Williams' project is not to show that any of the thought experiments gives us the *real* answer about the true criteria of PI. Rather, he tries to show that we have two conflicting intuitions about PI, one of which supports the 'mentalistic' thesis and the other one supports the 'bodily continuity' thesis of PI. The first set of thought experiments represents the first kind of intuition about PI while the second set of thought experiments represents the second kind of intuition about PI. And, because of the existence of two different kinds of intuitions about PI, mind-swapping sort of thought experiments based on memory exchange provide us conflicting results, and hence, are unsuccessful in final analysis. So, Williams' project fairly accommodates the alleged subject-dependency. Hence, this criticism is not appropriate to Williams' project.

So, Williams is successful in showing the alleged limitation of thought experiments regarding PI. The lesson of his discussion, as I have mentioned earlier, is that mind-swapping or body-swapping sort of thought experiments should not be considered as true guides in finding the criteria of PI. I am inclined to accept this lesson and would like to add here my observations that show that memory exchange, which is the central to all mind-swapping or body-swapping sort of thought experiments, is neither practically nor theoretically possible. If memory exchange is practically and theoretically impossible, then mind-swapping or body-swapping sort of thought experiments are mistaken because they are founded on memory exchange. Here I elaborate my observations:

(1) Memory exchange is practically impossible: In a slandered mind-swapping sort of thought experiment it is assumed that a person *A*'s memories, m_1, can be copied and pasted into another person *B*'s body; and *B*'s memories, m_2, can be copied and pasted into *A*'s body. It is also assumed there that after the memory exchange the *A-body-person* will do taking pre-experimental *B*'s memories as her *own* memories. And the same will be true of *A-body-person*. That is, a memory exchange is successful if the post-experimental *A-body-person* accepts m_2 as her own memories *now*; and the post-experimental *B-body-person* accepts m_1 as her own memories *now*. The whole experiment can be done in two ways: (i) the subjects of the experiments, e.g. persons *A* and *B*, are told about the memory exchange before the experiment takes place, and (ii) they are not told about the memory exchange before the experiment.

Now, suppose, the experiment is conducted following the first method, i.e. *A* and *B* are told about the memory exchange before the experiment

starts. Defenders of mind-swapping sort of thought experiments will claim that the post-experimental *A-body-person* will acknowledge m_2 as her own memories after the experiment is completed. But I think, things may go in a different way. Since pre-experimental *A* and *B* were told about the memory exchange, both of them will remember, after the experiment, that they have gone through a memory exchange sort of experiment. In that case, it is not unnatural that post-experimental *A-body-person* will claim that m_2 were her memories when she was *B-body-person*. And the post-experimental *B-body-person* will make a similar claim.[3] There is nothing irrational in such claims. Rather, these are the most natural and the most rational claims for post-experimental *A-body-person* and *B-body-person*. This feature shows that post-experimental *A-body-person* does not accepts m_2 as her own memories *now*. And the same is true of post-experimental *B-body-person*. That is, a true memory exchange has not been occurred.

Again, suppose, the experiment is conducted following the second method, i.e. *A* and *B* are not told about the memory exchange before the experiment is done. In that case, after awaking up, post-experimental *A-body-person* will observe that she *sees*[4] memory images, e.g. m_2, where she finds someone else instead of herself. She will be surprised and probably will see a psychologist with the complaint that she *remembers* something which is not her own memory. The same thing will happen to post-experimental *B-body-person*. That is, the post-experimental *A-body-person* does not recognize m_2 as her own memories *now*. And the same is true of post-experimental *B-body-person*. This feature of rejecting m_1 and m_2 as their memory by post-experimental *A-body-person* and *B-body-person* respectively proves that a true memory exchange has not been occurred.

Thus, we see, whatever method we follow, a memory exchange between two individual persons is practically impossible.

(2) Memory exchange is theoretically impossible: Memory is causally connected with perception. We remember events[5] that we previously perceived. Perception, again, is causally connected with sensation. Sense-data are the raw-materials of our perception. Without receiving sense-data no perception is possible.[6] But how do we receive sense-data? We receive sense-data through sense organs. Sense organs require the physical presence of a lively body. That is, a memory impression is a memory of past perception based on sense-data received by a lively body. Taking this view of causal links in mind we can define memory in the following way:

> *Def: m is a memory of a person A if, (i) A recalls m; (ii) m is true in the sense that it corresponds the actual event e (i.e. m is the image of e); (iii) A is physically present at the event e (as the receiver of the sense-data relevant to the event e).*

Thus, according to the definition of memory any sort of memory exchange between two persons is theoretically impossible[7] because after the so-called memory exchange both persons fail to satisfy the condition (iii) since then they have bodies which are different from the bodies with which they encountered the event they are now recalling. What they recall, after the said memory exchange, are not memories, but a kind of *pseudo memories*. So, a genuine memory exchange is theoretically impossible.

Finally, Bernard Williams has shown that there are significant limitations of mind-swapping sort of thought experiments that are constructed to support philosophers' accounts of PI. Accepting the main spirit of Williams' project, I have tried to show that thought experiments based on memory exchange are practically and theoretically impossible. The upshot of my discussion is that mind-swapping or body-swapping sort of thought experiments based on memory exchange cannot be a proper way of the study of PI since they involve practical and theoretical impossibility. Hence, any account of PI which involves such thought experiments are fundamentally mistaken.

References :

1. All page numbers in this paper, unless otherwise mentioned, refer to: Bernard Williams, "The Self and the Future" in *Personal Identity*, edited by John Perry, Berkeley, Los Angeles and London: University of California Press, 2008.

2. Both set of thought experiments are similar in the sense that both of them include mind-swapping based on memory exchange. They are just described in different terms and different styles only.

3. Their *real memories*, as post-experimental *A* and *B*, will begin shortly after the experiment is done.

4. Memories come to us as images (memory-images). Most of the time people 'see' their own image when they recall memory-images.

5. Some philosophers, such as John Perry, distinguish between 'event-memory' and 'memory-that.' Event-memories' are memories of events that we witnessed or in which we consciously took parts. 'Memory-that' are memories of something which we never witnessed. Perry's example of 'memory-that' is 'Most of us remember that Columbus discovered America in 1492.' (Perry, p. 144). I do not agree with such distinction. For me, we do not remember Columbus' discovering of America. What we remember is the 'information' about Columbus' discovering of America. And receiving information is an event in which we take part. So, all memories, this way or that way, are event-memories indeed.

6. It could be argued that illusions and hallucinations are perceptions that do not involve exact sense-data. But the fact is that illusion is an admixture of sense-data and imagination whereas hallucination is formed by imagination only. But imagination itself depends on previous perceptions based on sense-data. Actually one cannot imagine anything about which one did not have any sort of previous perception and relevant sense-data.

7. Theoretical impossibility implies practical impossibility, but practical impossibility does not imply theoretical impossibility. It is practically impossible to survive in a temperature of 5000C. But it is theoretically possible; we can imagine that scientists invent a shell or outer-covering that protects us from extreme temperature. But we cannot imagine drawing a 'round-shaped triangle' because it is theoretically impossible; the very definition of triangle ensures that it cannot be round-shaped by any means. So, *theoretical impossibility involves definition*. Something is theoretically impossible if something is impossible by definition. Practical impossibility and theoretical impossibility can be called material impossibility and formal impossibility respectively.