Personal Identity and False Memories

Danil Razeev [0000-0002-5129-7532]

Institute of Philosophy, Saint Petersburg State University, Russia
d.razeev@spbu.ru

Abstract. In current philosophy of mind, there are two main approaches to the question of personal identity. The first one claims that personal identity is based on our memory and for several decades has been known as a psychological approach to the problem. The second one has been called an animalistic approach and considers personal identity as a biological property of human beings or as a specific feature of our bodily continuity. The experiment on creating false memories in mice brains, recently conducted at Massachusetts Institute of Technology (MIT), seems to shed new light on the question of personal identity, taking into account the fact that the mouse brain is morphologically quite similar to our brain. The purpose of my paper is to consider whether the above-mentioned experiment supports one of the approaches: the psychological or the animalistic. Using the conceptual instrumentarium of contemporary analytic philosophy and cognitive phenomenology, I differentiate between strong and weak false memories and I argue that we cannot consider the conducted experiment to have created false memories in the strong sense. I develop a thought experiment showing what it would be like to experience an implanted (weak) false memory in the human brain. I conclude that there is not and cannot be an experience of the (strong) false memory.

Keywords: Philosophy of cognition, Personal identity, False memory, Animalism.

1 Introduction

Could a computer or robot be a person? Contemporary philosophers and scientists do not offer an unequivocal answer to this question. Some of them claim that being a person is inherent only to highly developed biological organisms, particularly humans, and cannot be found in non-biological matter. Others express some optimism in this regard and claim that, in the future, it will be possible not only to create artificial intelligence possessing genuine personality but also to transfer a biologically-based personality to an artificially-built one and vice versa.

The problem of personal identity is deep rooted in the history of philosophy. The British philosopher John Locke has been one of the most influential figures in discussing the question about personal identity. In his famous book "An Essay Concerning Human Understanding" [1] he points out that we cannot find the only unitary criterion of identity for all that exists. Locke's argumentation, let me present it in a

slightly modern and free manner, suggests dividing identity into three basic types: the first one is the identity of a thing, the second is the identity of a living organism and the last one is the identity of a person. The identity of a thing depends on the identity of material stuff, the identity of a living body goes back to its persistence as an organism, i.e. as a unified whole, and the identity of a person has its roots in the capacity to maintain a kind of self-representation through time. Let us give some examples of Locke's typology of identities. Identity of a thing does not allow us to identify the statue of David with the huge piece of marble, from which Michelangelo's masterpiece was sculptured, rather we deal with a process whereby one thing became something else. An example of the second type of identity would be a cat having been identified by its owner as the same living organism through time, although in its old age it does not look like that pretty kitten, which first entered the house. More confusing is the situation with personal identity and each of us, human animals, could count as an example for this kind of identity. Intuitively, we understand the difference between the identity of a living body and personal identity, as in the case of a human being falling into a vegetative state, where the body remains the same, but its carrier loses his bodily citizenship so to speak. Many contemporary philosophers think that the question about personal identity should be considered beyond the framework of our living bodies.

Even careful usage of the term identity can lead us to a set of very difficult questions, such as: 1. Does the identity of a thing mean the material identity atom by atom? 2. Where can a material boundary for a living thing be found? 3. Can a digital copy of a person be created? 4. Would a person remain the same, if they were reproduced using a different material carrier? 5. What would happen to a person if they were reproduced using two carriers, materially identical atom by atom? And so on.

2 Experimental data

The question about what makes us identical through time has not found any unambiguous answer and has been discussed by many philosophers studying the problem of personal identity since Locke's time. Nowadays, there exist two general approaches to the problem: the psychological one, which is sometimes called psychological reductionism [2, 3, 4, 5], and the somatic one, the so-called animalism [6, 7, 8, 9]. According to the first approach, a criterion for personal identity has to be found in our psychological continuity over time. The second one tries to find this criterion in the persistence of our bodily organization. Psychological reductionism has been continuing Locke's attempt to find a certain criterion for personal identity in the mechanism of memory. Animalism regards this approach as conceptually wrong and claims that the identity of a person can be completely reduced to the identity of this or that living body, in our case to the identity of a living body of the human type or, in short, of a human animal.

In my paper I would like to consider in detail a very interesting scientific experiment, recently conducted by neuroscientists at Massachusetts Institute of Technology (MIT). It pretends to change our understanding of genesis and structure of subjectivity

and shed new light on the contemporary discussion about whether or not personal identity could count as a special and irreducible type of identity. The neuroscientists Susumi Tonegawa and his colleagues at MIT claim to have created a false memory in the brain of a living organism [10].

First of all, let me recall the details of the experiment on the mouse brain. For the experiment scientists used genetically modified mice whose neuronal activity in hippocampus, a specific region in the brain responsible for memory, could be activated or deactivated by flashes of light, using a special laser device attached to the mouse brain.

At first, the mice were placed in a box (box A) with a comfortable environment and the neuroscientists were able to trace the neuronal activity in the mice's hippocampus. After that, the mice were moved to another box (box B), where their memories about being in Box A were activated with a laser while at the same time their feet were shocked with electricity. Using this technique in the mice brains, an association of being in Box B and experiencing some fear there was created. Being placed again in Box A, the mice behaved as if they remembered some negative experience in Box B that in reality had never happened. In such a way the neuroscientists came to a general conclusion that they had created in the mice brains a false memory or a memory about something that never actually happened.

Although the experiment was conducted on mice brains, in my opinion, it has very serious philosophical and ethical consequences for the understanding of our own subjectivity. Even though morphologically the mouse brain and the human brain are similar, to draw conclusions about the structure of our subjectivity based only on the results of the experiment would not be correct. Nevertheless, in philosophy we can conduct so-called thought experiments. As philosophers we are permitted to suggest that a set of neural events that happened in the mouse brain during the experiment could have happened in the human brain, despite the fact that it cannot be verified at the present moment due to the lack of technology or because of ethical restrictions. In other words, I would like to discuss some philosophical consequences that would have arisen if the experiment had been conducted on humans and it had resulted in the creation in the human brain of false memories about some events that in reality never happened.

If we extrapolated this experiment from mice to people, not taking into account technical and ethical aspects [11], it could work in the following way. At first, a volunteer is placed in a blue room with a comfortable environment. Then, s/he is moved to a red room with an uncomfortable environment, where neuroscientists, using a special technology, would activate a specific region in the volunteer's brain responsible for the memory of his/her previous presence in the blue room and thereby create an additional association between his/her presence in the blue room and the uncomfortable environment in the red room. Lastly, s/he is moved back to the blue room. The result of the experiment is expected to be as follows. The volunteer will remember some negative experience about her/his previous presence in the blue room that in reality never happened to her/him.

Having become acquainted with the details of the experiment, I would like to involve it in the contemporary discussion about personal identity. Could the experiment shed new light on the problem of personal identity?

3 Evaluation and Discussion

The experiment seems to challenge the fundamental role that memory, according to the psychological approach, plays in the constitution of personal identity. Obviously, memory, being controlled from the outside, cannot be an intrinsic feature of personal identity and should be regarded rather as an extrinsic, additional mechanism. If, using a specific technology, neuroscientists were able to switch on or off some neuronal populations in the hippocampus, and to make someone recall events that never happened or make them suppress others that did, and, thereby, were able to manipulate their behavior in the future, personal identity would lose its central and fundamental mechanism, the mechanism of memory, which would disintegrate into pure animalistic mechanisms. It would mean that we do not need a personal identity in order to exist and we are nothing but highly organized animatons possessing several sophisticated cognitive mechanisms, one of which is memory [12, 13].

At a first glance, the experiment seems to support the animalistic approach. Nevertheless, I suggest being cautious and raising some important questions concerning the experiment before drawing radical philosophical conclusions. The first question concerns the status of false memory in the experiment. More specifically, with what kind of memory are we dealing in the experiment: is it false memory or rather a kind of modified memory? The second question is related to the very process of memorization in the experiment. I am asking, whether actually experiencing an event is a necessary condition for creating memory about it or whether memory can be created even if the event was not actually experienced. The third question is about the role that the mechanism of memory seems to play with regard to the identity of a person and whether modifying memory or creating false memory can significantly change personal identity.

The question about the status of memory in the experiment on the mouse brain is a core one. Let me differentiate between creating a new memory and modifying the already existing memory. Neuroscientists called the type of memory they dealt with in the experiment “false memory”. I think they misinterpreted the very concept of false memory. In my opinion, the experiment dealt with what I would call distorted memory. Let us take a look at the details of the experiment from this perspective. Firstly, being in Box A, the mice stored some information about being in a positive environment. Secondly, being moved to Box B, their memories about being in Box A were activated and additionally associated with experiencing some fear, because their feet received an electric shock. Thirdly, being placed again in Box A, the mice recalled a distorted memory, i.e. their original memory of being in Box A was superimposed by their memory of being in Box B. I think that the overlapping of memories can lead to creating distorted memory and is part of our everyday psychological process. We only need to consider cases of eyewitnesses at a crime scene. Being asked,

just after an incident they usually cannot recall any specific information, but later, step by step, they begin to recall details of the incident very vividly. In a broad sense, all our memories can be regarded as distorted memories. Individual pieces of memory do not exist in isolation. Each one is always recalled in a new context and, being recalled in the present, has already been modified. In my opinion, in the experiment scientists did not create in the mice's brain false memories in a strong sense. They mixed together the memories that had existed in the mice's brain before, resulting in what we can call distorted memory or, probably, false memory in a weak sense.

Creating false memories in the strong sense is connected to the second of the above-mentioned questions, namely, whether it is possible to create such a memory that would not refer to an event experienced earlier by a subject [14]. If we tried to imagine the conditions of such an experiment on false memories in the strong sense, they would probably be as follows. Subject number one is placed in a blue room and subject two in a red room. Then both subjects are moved to a green room where, using a new sophisticated technology, neuroscientists exchange the subjects' memories about being in the blue and red rooms respectively. After that, the subjects are put in the previous rooms, but now in reverse order: subject number one in the red room and subject two in the blue one. The experiment would succeed if subjects could report remembering already having been in the rooms. It would seemingly prove that neuroscientists could create false memories in subjects about their being where they had never been before. Nevertheless, I am afraid that even this hypothetical experiment, were it technically possible in the future, would not prove that false memories in the strong sense are possible. The problem is more difficult then it appears to be. Even if subjects could recall each other's memories or have their own memories exchanged, it would not mean that false memories in the strong sense had been created in their minds. In my opinion, even in this case, each subject's memories would remain false memories in the weak sense, because each of them would, in the end, refer to an experience undergone earlier by another subject. In order to be recalled, false memories must refer to experiences that have actually happened before, regardless of which of the subjects' minds was involved. I am afraid that false memory in the strong sense is not possible in principle, i. e. logically not possible. In other words, only something that has happened before in the actual experience of a conscious mind can be recalled in our memory. Whether experiencing being and recalling being have to be the same thing is another question.

4 Conclusion

To what extent is the mechanism of memory crucial for personal identity? To what extent being a person presupposes recalling in memory what happened in your own experience and not in another's experience? Often we coordinate our behavior by recalling in our semantic memory something that never happened personally to us. I do not need to be hit by a car in order to realize that the street should be crossed on a green light. Other animals certainly possess a similar mechanism of memory. They,

like us, are capable of learning something by recalling the experience of others. And they can do it without being persons.

Could the results of the experiment on false memories be used as empirical support of the animalistic approach to personal identity and contra the psychological one? I don't think so. All that the experiment has proven is the existence of a certain type of memory that can be activated bypassing the phase of actual experience. It means that as human animals we are able to maintain our life, process information cognitively, regulate behavior and we can carry out all of these activities using a specific type of memory, which is not accompanied by awareness. At the same time, it does not prove that we do not possess a different type of memory, which defines us as persons and takes us beyond mere human animals. It means that the mechanism of memory still remains a vessel for personal identity.

In general, our analysis of the experiment on false memories shows the following. Firstly, the experiment does not support the animalistic approach to personal identity. If animalism were true, we could not exist except as animals and personal identity would be a phenomenon that is solely inherent to some highly developed biological organisms. Secondly, our analysis emphasises support of the psychological approach to personal identity. If the psychological approach is correct, then memory is a core mechanism of personal identity. The psychological approach gives us hope for the development, in the future, of such a form of artificial intelligence that could possibly possess a genuine personality.

Acknowledgments. This work was supported by the Russian Foundation for Basic Research under Research Grants 18-011-00840.

References

1. Locke, J.: An Essay Concerning Human Understanding. Oxford University Press, Oxford (1975).
2. Parfit, D.: Reasons and Persons. Oxford University Press, Oxford (1984).
3. Shoemaker, S.: Self-Knowledge and Self-Identity. Cornell University Press, Ithaca (1963).
4. Shoemaker S., Swinburne R.: Personal Identity. Blackwell, London (1984).
5. Unger P.: Identity, Consciousness, and Value. Oxford University Press, Oxford (1990).
6. Hudson H.: A Materialist Metaphysics of the Human Person. Cornell University Press, Ithaca (2001).
7. Olson E.: The Human Animal: Personal Identity Without Psychology. Oxford University Press, Oxford (1997).
8. Olson E.: What Are We? A Study in Personal Ontology. Oxford University Press, Oxford (2007).
9. Thomson J.: People and Their Bodies. In: Dancy, J. (ed.): Reading Parfit, pp. 202–209. Blackwell, London (1997).
10. Ramirez, S., Liu, X., Lin, P., Suh, J., Pignatelli, M., Redondo, R., Ryan, T., Tonegawa S.: Creating a False Memory in the Hippocampus. Science, 341(6144): 387–391 (2013).
11. Liao, S: The Ethics of Memory Modification. In: Bernecker, S., Michaelian, K. (eds.) Routledge Handbook of Memory, pp. 373–382. Routledge, New York (2017).
12. Shettleworth S.: Cognition, Evolution, and Behavior. Oxford University Press, New York (2010).

13. Millin, P., Riccio, D.: False Memory in Nonhuman Animals. Learning & Memory, 26 (10): 381–386 (2019).
14. Vetere, G., Tran, L., Moberg, S., Steadman, P., Restivo, L., Morrison, F., Ressler, K., Josselyn, S., Frankland, P.: Memory Formation in the Absence of Experience. Nature Neuroscience, 22(6): 933 (2019).