On Hume's Theory Of Self[1]

Winner of the 1996 Larry Taylor Award

Yiwei Zheng

One of the most frequently discussed topics in Hume's Treatise is his theory of self. But few have examined in detail the relationship between his theory of self and his identity theory in general. In this paper I wish to show that Hume's theory of self not only suffers from external difficulties, coming from outside of his system: i.e., the familiar Kantian objections, but also suffers from an internal difficulty, that is, his theory of self cannot be rendered consistent with his identity theory in general.[2]

One common criticism of Hume's theory of self is that our idea of self cannot be a fictional product of the association of ideas, since it is presupposed in explaining the associative principles.[3] Another common criticism also claims that the idea of self cannot be a fictional product, but the reason is simply that the general idea of explaining the identity of self by means of the association of ideas is mistaken. Without going deeply into these criticisms, we find that by claiming that the idea of self is not a fictional product of the association of ideas, both criticisms do not merely aim at Hume's theory of self in particular, rather, they both challenge the underlying rationale of Hume's identity theory – that is, to account for the identity of objects on the basis of the association of ideas, and thus, they both challenge Hume's identity theory in general. However, whether or not one can make a good case out of that is not the line that I will pursue here.

My aim is more humble. I do not try to challenge Hume's identity theory in general, and it is not my concern to demonstrate whether or not Hume could give a plausible explanation of the identity of external objects. What I try to prove is that Hume's theory of self does not follow from his identity theory in general, that is, even if his identity theory in general is sound, and even if he could give a plausible explanation of the identity of external objects, this cannot be translated to the idea of self.

In *Book I* of the *Treatise* Hume introduces three associative principles that constitute the major framework for his identity theory. They are the resemblance principle, the contiguity principle, and the causation principle. Since these are the only principles Hume gives in the *Treatise* to account for the identity of objects, "it follows, that our notions of personal identity, proceed entirely from the smooth and uninterrupted progress of the thought along a train of connected ideas, according to the principles above-explained (the resemblance, contiguity, and causation principles)."[4] But do these principles work well in explaining the fiction of the

idea of self? Let's look at them closely.

First, let's examine whether the resemblance principle could explain the idea of self. Consider two perceptions, a perception of a table and a perception of a chimney. Both are my perceptions, but do they resemble each other? At first sight they do not seem to, since a table is obviously different from a chimney. However, the situation might change if we think a bit further.

One move that might make the resemblance principle work is to dissect a perception into a combination of an objective side and a subjective side. As Geoffrey Scarre says, "Objectively, a perception may be of a table or of a chimney; subjectively, it might be Hume's perception of a table or mine of a chimney."[5] And thus, a sequence of perceptions presents "from the objective angle, a plurality of distinct existences, while yet constituting, from the subjective point of view, a unitary self."[6] Accordingly, though the perception of a table has nothing similar to the perception of a chimney on the objective side, they do resemble each other on the subjective side, which might account for the idea of self. Furthermore, by claiming that these two perceptions are similar on the subjective side, we are not saying that they really have something in common on the subjective side – i.e., the character of belonging to a unitary self. It is true that we might use the idea of a unitary self to further explain the resemblance of the two perceptions on the subjective side, but this move does not seem to be inevitable. Thus, here we do not have to worry about the kind of circularity Barry Stroud proposes, that to explain the idea of self, the resemblance principle is appealed to; and to explain the resemblance principle, the idea of a unitary self is appealed to.[7]

It should be noted that the above move can be taken in two different ways: 1. it might be taken to mean that a perception *by itself* has both a subjective side and an objective side; 2. it might be taken to mean that while not having a subjective side and an objective side by itself, a perception can be *viewed* both subjectively and objectively. Let's consider them respectively.

With respect to "2," it does not seem to be compatible with Hume's theory. By claiming that a perception can be *viewed* both subjectively and objectively, we implicitly reckon and give priority to the Kantian transcendental consciousness, since our claim entails a fundamental shift from perceptions as what they are to the constitutive roles of the transcendental consciousness. Needless to say, this move would fundamentally affect Hume's system as a whole.

Next, let's examine "1." First, anyone who endorses the intentionality thesis[8] cannot accept "1."[9] For by claiming that a perception has a subjective side, don't we create an opaque residuum inside the perception, as if it were not wholly exhausted by its object, as if it contained something more than the perceived object, solid and unpenetratable? But how can this be possible? According to the intentionality thesis, even a simple observation would refute it: when we perceive a

table, in that perception only the object "table" is presented, and besides that, nothing else. In brief, the subjective content of a perception is regarded as a myth for the believers of the intentionality thesis. Second, anyone who holds the doctrine that consciousness is consciousness through and through– consciousness does not contain any unconscious pert – cannot accept "1." For if a perception really contains a subjective part, and if we are conscious of the subjective part all the time when we have the perception, how can we overcome the subjective barrier to reach the object, how can we explain or explain away the fact that the object in the perception appears to us as being objective, and how do we understand the process in which an "objective thing" is created out of the subjective data? Here again, we seem to encounter an insuperable impasse. Third, even if one does not accept the intentionality thesis or the doctrine that consciousness is totally translucent, "1" still has serious problems. One difficulty is concerning the identity of a perception. Consider again the perception of a table. According to the resemblance principle, to account for the identity of the table, the objective part of the perception is utilized; and to account for self, the subjective part of the perception is utilized. Now, the problem is this: how can we know that the perceptions we have in the two cases are really the *same* perception, rather than two different perceptions, if in the former case only the objective part is presented and utilized,[10] while in the latter case only the subjective part is presented and utilized? Furthermore, how do we understand the relation between the objective part and the subjective part in a single perception? What binds the two parts together? It seems that there must be a necessary connexion between the subjective part and the objective part, otherwise the perception itself would become a fictional object, created on the basis of the three associative principles. However, it seems to me quite doubtful that there is any room within Hume's theory for this necessary connexion, since this necessary connexion is obviously not a logical connexion. In sum, without giving a satisfactory solution to the above problems, it seems to me implausible to attribute a subjective aspect to perceptions.[11]

We have seen that there are serious difficulties in explaining the idea of self on the basis of the resemblance principle. The situation does not seem to be better when we try to explain self on the basis of the contiguity principle, since the previous arguments can be easily translated to the contiguity principle, which due to the limitation of space I shall not repeat. In the following I will focus on the causation principle. In particular, I will examine whether the fiction of the idea of self can be explained by the causal relations among perceptions.

What is the nature of the causation principle? What do we mean by saying that there is a causal relation between a perception A and a perception B? Certainly the causal relation is not a "real connexion" between the perception A and the perception B, but a natural gentle force of associating A and B on the basis of habit. In

other words, to say that there is a causal relation between A and B, the sequence <A, B> (or <B, A>) must have appeared many times in the past, such that in the future whenever I have the perception A (or B), I naturally introduce the perception B (or A). And it is this natural force of thoughts to associate A and B on the basis of the past repetitious experience of the sequence <A, B> (or <B, A>) that accounts for the causal relation between A and B. Here the key point is that in order to associate A and B, the sequence <A, B> (or <B, A>) must have been constantly perceived in the past without violation. However, if the sequence <A, B> (or <B, A>) is just perceived for only one time or few times, it is then insufficient to attribute the causal relation to A and B. This is a rather simple point, but we shall see that it yields intolerable results for the causal account of self.

Suppose the causal account of self works. Suppose $\{T_1, T_2, ..., T_n\}$ represents a series of times from my birth to death. And suppose T_1 represents the time of my birth, T_n represents the last moment that I am conscious, and T_m represents an earlier time than T_{m+1} ($1<=m<=N$). Let P_m be the set of all perceptions (we have at T_m) which are or can be used to create the idea of self at T_m, if I had already got the idea of self by T_m and let P_m be the empty set if otherwise.[12] Since during my life I do have the idea of self, in the series $\{T_1, ..., T_n\}$ there must be a time T_a ($1<a<m$)[13] such that at T_2 the idea of self is accessible to me and P_a is not empty. Given all these, if follows that for any b>a and b<-n, P_b must be a subset of P_a. Let's prove it.

To show that P_b is a subset of P_a, we have to first prove that P_{a+1} is a subset of P_a. Let's try an indirect proof. Suppose P_{a+1} is not a subset of P_a. Then P_{a+1} must have at least one member X which does not belong to P_a. Yet since the perception X is not contained in P_a, it does not participate in any causal chain in P_a. Furthermore, X's mere appearance with others in P_{a+1} is insufficient to constitute any new causal relation between X and some other perceptions, because the logic of causal relations demands the repetitious experience of a certain sequence of relevant perceptions, which we do not have here.[14] Thus, X cannot be a member of any causal chain in P_{a+1}, which means that it cannot be a perception which is or can be used to create the idea of self at T_{a+1}. According to the definition of P_m, we have that X is not a member of P_{a+1}, which contradicts the given fact that X is a member of P_{a+1}. Therefore, our assumption is false, and P_{a+1} is a subset of P_a.

Following the previous line, we can also prove that P_{a+2} is a subset of P_{a+1}, P_{a+3} a subset of P_{a+2}, and so forth. To generalize, we get the result that for any b>a, P_b, on the basis of the assumption that the causal theory of self works.

But this thesis is absurd, since it says that after a certain time I will never have any new perception(s) which can be viewed as my perception(s), and this is completely at odds with common sense! Consider my near-death experience for example,[15] it is certainly my perception, and thus belongs to P_n.[16] Furthermore, since my near-death experience can only be had once in my life (in normal situations), it

does not belong to P_a. Hence, Hume's causal account of self does not work.

To sum up, we have shown that Hume's three associative principles fail to explain the identity of self. To make his theory of self tenable, Hume has to either appeal to some other associative principle(s) to explain the fiction of the idea of self, or demonstrate that there is a "real connexion" among different perceptions used to produce the idea of self.[17] But since the above three principles are the only ones Hume gives in the *Treatise* to account for the identity of objects, we conclude that there is an unsolvable inconsistency between Hume's theory of self and his general theory of identity and association of ideas.

Notes

1. I wish to thank Prof. Jane L. McIntyre at Cleveland State University for stimulating my interest in this subject. I also wish to thank Prof. Gary Cesarz at the University of New Mexico for his helpful comments on this paper.

2. It might be argued that this difficulty is indeed what Hume has in mind when he criticizes his theory of self in the *Appendix of the Treatise,* and there is some textual evidence supporting this claim (see David Hume, *A Treatise of Human Nature,* ed. L.A. Selby-Bigge (London: Oxford University Press, 1965), 635-636). But this is not my concern in this paper.

3. For example, Barry Stroud thinks that the idea of a unitary self is presupposed in explaining causality. See Barry Stroud, *Hume* (Routledge and Kegan Paul, 1977), 135. See also S.C. Patten, "Hume's Bundles, Self-Consciousness and Kant," *Hume Studies* (1976), 59-75.

4. Hume, *Treatise, 260.*

5. *Geoffrey Scarre, "What Was Hume's Worry About Personal Identity?"* Analysis (1983), 219.

6. Scarre, *ibid.*, 219.

7. This is not exactly Stroud's point. In his book Stroud mainly worries about the circularity involved in explaining causality. But the structure of the circularity is the same. See Stroud, 135.

8. Here I mean the thesis (as proposed by Brentano, Husserl and Sartre) that consciousness is always consciousness of something.

9. However, the choice of the intentionality theory is not recommended in the context of this paper, since it is incompatible with Hume's identity theory in general, and thus trivializes my arguments.

10. If the subjective part is also presented, as I said before, how can we understand and explain the process in which an "objective thing" is created out of the subjective data?

11. Another way of defending the distinction of the subjective and objective sides of a perception might be the following: Consider Sartre's distinction of the positional and non-positional consciousness. According to Sartre, every consciousness is positionally aware of its object, and non-positionally aware

(of) its conscious act. So we might let the subjective side of a perception be the Sartrean non-positional consciousness, and the objective side be the positional consciousness. However, in Hume's theory there cannot be any space for the Sartrean non-positional consciousness, since the Sartrean non-positional consciousness is the peculiar way that consciousness is aware (of) itself, and to accept that would lead to the result that the idea of self is not a fiction (since every perception has the same structure of the non-positional self-consciousness), which is intolerable for Hume.

12. This definition implies that for any b>a, P_b must be non-empty if P_a is non-empty. This is consistent with common sense, since we know that if at certain time I have the idea of self, I will continue to have it afterwards (certainly this rules out the cases of mental illness and other abnormal situations).

13. This is the case because I didn't have the idea of self at the time of my birth, and I do not wait until the last moment of my life to create the idea of self.

14. Here I assume that T_{a+1} represents the first time that we have the perception X. But my argument does not depend on this assumption. Even if T_{a+1} does not represent the first time that we perceive X, we can always identify the first time that we perceive X, and change the value of T_{a+1}, making it represent that time. Then the same argument follows.

15. It might be thought that the example of near-death experience alone is able to refute the causal theory of self, and we do not need to go through all previous proofs. One might be able to make a strong case out of that, but in the text I intend to prove something stronger, that is, even if all experience we have is repetitious experience, the causal account of self still fails to work.

16. Otherwise how can I know that I am going to die?

17. Hume seems to realize this in the *Appendix* of the *Treatise*. See Hume, *Treatise,* 636.