

*How to determine which is the true
theory of personal identity*

Richard Swinburne

THE PROBLEM

The simple view of diachronic personal identity holds that personal identity is not constituted by continuities of mental or physical properties or of the physical stuff (that is, the bodily matter) of which they are made, but is a separate feature of the world from any of the former, although of course it is compatible with personal identity being caused by such continuities. On the simple view, as I shall understand it, a person P_2 at t_2 can be the same person as a person P_1 , at an earlier time t_1 , whatever the physical or mental properties and whatever the body possessed by each person. P_2 may not at t_2 remember¹ anything done or experienced by P_1 at t_1 or earlier, and may have an entirely different character and a totally different body (including brain) from P_1 at t_1 . The main arguments in favor of the simple view consist in adducing thought experiments in which persons undergo radical changes of mental life and bodily constitution, and in which – it is claimed – it is “possible” that they continue to exist; from which it follows that continuities of the kind mentioned are not necessary for personal identity.

I begin with one example (of very many which have been put forward) to indicate the role of thought experiments in supporting the simple view. Suppose I have a severe brain disease affecting the right brain hemisphere. The only way to keep my body functioning is to replace this hemisphere. So the doctors remove my current right hemisphere and replace it by a right hemisphere taken from a clone of me. The new right hemisphere, let us suppose, contains the brain correlates of (that is the neurons, the states of which are the immediate causes of) similar but slightly different memories and character traits from mine. The resulting person would presumably to

¹ Ordinary language sometimes assumes that only true beliefs are correctly called “memories.” Thus it assumes that if I am correctly said to have a “memory” that I did so-and-so, I really did so-and-so. I shall not follow that usage here, but shall understand by “memory” what on that usage would only be an apparent memory: it seeming to the subject that he “remembered” so-and-so.

some extent behave like me and remember having done what I did and also to some extent behave like my clone and remember having done what my clone did (when what I did was different from what my clone did). Now suppose that the disease spreads to the left hemisphere, and that too – two years later – is replaced by a left hemisphere taken from a different clone of me, again containing the brain correlates of similar but slightly different memories and character from mine. Then my body would be directed by a brain made of totally different matter and sustaining rather different memories and character from those I had two years previously. Yet presumably to some extent, but to a lesser extent than after the first operation, the resulting person would still behave like me and remember having done what I did.

But would the resulting person be me? That person would be a person largely continuous with the earlier me two years ago, apart from having had two large brain operations. One might think that the continuity of many mental and physical properties over this period has the consequence that the same person continues to exist. Yet the resulting person would have none of the brain matter and only some of the memories and character which were previously mine. I suggest that it is totally unobvious whether in this situation the resulting person would or would not be me. Yet the question “Would the resulting person be me?” is logically equivalent to the question “Would I survive the operations?” and so have the (pleasurable or unpleasant) experiences of the resulting person. And surely no one about to have a serious operation can think that the question of whether he will “survive” a brain operation is simply a question requiring an arbitrary decision about which of two senses we should give to the word “survive.” We (or at least most of us) seem to understand the alternatives as mutually incompatible factual alternatives – that I survive, or that I do not survive – in one clear and natural sense of “survive.” Yet it is totally unobvious what is the answer. If you think that – one way or other – the answer is obvious, it is easy to alter the thought experiment in such a way that the answer is no longer obvious. If you think it is obvious that the continuity of at least half the brain matter over each of the operations two years apart insures that I continue to exist, suppose the second operation to be performed after only one year or six months. If you think it obvious that when half my brain matter is removed in one operation I no longer exist, suppose a series of operations replacing only a tenth of the matter each time.

In such a situation, which I call an ambiguous situation, it does seem possible that I have survived (i.e. continued to exist), and possible that I have not survived; and yet that we do not know (and have no further means of finding out) whether I have or have not survived. If what seems possible is

indeed possible, my survival does not require any particular degree of strong physical and mental continuities² which make it obvious that I do survive. It then follows that the difference between the situations of different degrees of continuity consists in the strength of the evidence that I continue to exist. Under normal conditions of very strong continuity of body (and in particular of the brain, the physical sustainer of mental life), memory (of what happened to a person with that body) and character, it is enormously probable that I continue to exist; it becomes less and less probable until we reach the ambiguous situation where it is as probable as not that I continue to exist.

Why it is enormously probable that under those normal conditions I continue to exist is first because it is a fundamental epistemological principle that (apparent) memory beliefs are probably true (in the absence of counter-evidence), and my personal memories (that is memories “from the inside” about what I did and experienced) concern the actions and experiences of the person who had a brain strongly continuous with my present brain. Unless memories as such (in the absence of counter-evidence) are probably true (and so do not require to be rendered probable by evidence of some other kind in order to be probably true), we would know very little about the world. For we depend on memory for all the knowledge which we believe that we have acquired from others about history and geography, etc.; and while my memory of these things may coincide with yours, at any time I depend on my own memory of what others have told me for my belief that our memories do coincide, and so the personal memory of each of us must as such be probably true if we are to have virtually any knowledge at all. And the second reason why it is enormously probable that under those normal conditions I continue to exist is that the simplest, and so most probable, hypothesis supported by the strong continuity of memory and character sustained by the same brain is that these are mental properties belonging to the same person. It would be less simple, and so less probably true, to suppose that the memory and character strongly continuous with the previous memory and character sustained by a brain having strong continuity with the previous brain are the memory and character of a different person. So being the same person does not entail strong continuity of brain, character and memory; but the latter is good evidence of the former. This is the simple view.

² By brain, memory or character being “strongly continuous” with a previous brain, memory or character, I understand that there exist between them both what Derek Parfit (1984, ch. 10) calls strong “connectedness” (that is strong similarity) and what he calls strong “continuity” (that is overlapping chains of strong connectedness), the continuity of memory and character being causally sustained by the strong continuity of the brain.

Some philosophers hold that personal identity, like the identity of artifacts, can be a matter of degree. On this view a later person can be only partly identical to some earlier person, and so the result of such operations as I have described might be that the resultant person was only partly me. I do not myself think that it is logically possible that some person be partly me. But even if this were a possible result of the operations, it does not seem to be a necessary truth that the operations will have this result, because the history of all the physical bits and all the mental properties associated with them seems compatible with the subsequent person not being partly me. It still seems possible that, just as the resulting person is fully me if I have both heart and liver replaced, so after the half-brain transplants the resulting person is still fully me; and it is also possible that it is not me at all. Yet if we include the subsequent person being partly me as a possible result of the operations, we would now be ignorant about which out of three (rather than two) possible results of the operations had in fact occurred. If what “seems possible” is possible, that I survive the operations not merely in part but wholly (or alternatively not at all), partial survival is compatible with the simple view.

The alternative to the simple view, the complex view, claims that personal identity is constituted (not merely evidenced) by a certain particular degree of continuity of physical and mental properties and of the matter which forms a person's body. The main arguments in favor of this view are that the paradigm examples of personal identity are all ones in which there is continuity of these kinds, and that the simple view leads to contradictions. There are innumerable varieties of the complex view according to which degrees of continuity ensure the identity of a later person with an earlier person. One variety is the view that the concept of personal identity has no application outside normal situations of strong physical and mental continuities. Another variety of the complex view holds that necessarily (not merely possibly, as in the version of the simple view) personal identity is a matter of degree. The weaker the continuities, the lesser the degree to which the later person is the same as the earlier person. Again there is an issue about this variety of the complex view, as about the similar variety of the simple view, as to whether the notion of partial identity of persons makes any sense.

LOGICAL POSSIBILITY

Which of these views is correct depends on what is possible, and so I come to a central topic of this chapter: what is it for some situation to be “possible,” and how can we know whether it is or not? We can have no discussable

knowledge of possibilities (or necessities or impossibilities) which cannot be expressed in sentences, and so I will discuss the question of which situations are “possible” (or whatever) by discussing which sentences describing situations describe “possible” (or whatever) situations, and I shall call such sentences “possible” (or whatever) sentences. In talking about possibility in this kind of context, we are talking about the widest or weakest kind of possibility there can be, which I shall follow most contemporary philosophers in calling “metaphysical possibility.” Some state of affairs may not be practically or physically possible, but it may still be metaphysically possible. Hence metaphysical necessity and metaphysical impossibility are the narrowest or strongest kinds of necessity or impossibility. Among metaphysical possibilities, etc., are ones discoverable *a priori*: that is, discoverable by reflection on what is involved in the claim made by the sentence. I shall call these logical possibilities, etc. No sentence could be more strongly impossible than a self-contradictory sentence. It claims both that something is so and also that it is not so (and is normally expressed by a sentence of the form “*s* and not-*s*”). For such a sentence could only be true if that something is so, and the sentence asserts that it is not so. But any sentence which entails a self-contradiction would be as strongly impossible as a self-contradiction. For similar reasons no sentence could be more strongly necessary than one whose negation entails a self-contradiction. Such necessities and impossibilities are logical ones, since they are discoverable by deriving the relevant contradiction by means of the rules of the language.

I see no reason to suppose that there are any other *a priori* impossibilities as strong as those which entail a contradiction, or any other *a priori* necessities as strong as those whose negation entails a contradiction. To begin with the case of impossibilities: what is asserted could only be *a priori* impossible if the impossibility is detectable merely by understanding what is involved in what is actually said. To be impossible a sentence must have the form of a declarative sentence in which the component words already have a sense in the language. It will be a subject–predicate sentence, an existential generalization, or some other one of many recognized forms of declarative sentences. It will – to put the point loosely – assert something about some substance or property or event or whatever that it has or does not have some property or relation to some other substance, property, etc.; or that there are or are not certain substances, properties or whatever. Words have a sense insofar as it is clear what are the criteria for an object, property or whatever picked out by the word being that object, property or whatever. They therefore delimit a boundary to the sort of object or property it can be or the sort of properties

it can have. Hence it will be inconsistent to affirm that an object picked out by some word is of a kind ruled out by the very criteria for being that object. And the form of any sentence will exclude some alternative; and so it will be inconsistent to affirm the sentence together with that alternative. If a sentence is not inconsistent in these ways (or does not entail one that is), we would have no reason to suppose that that sentence is *a priori* impossible. A similar argument shows that we would have no reason to believe that any sentence whose negation is not inconsistent in these ways is *a priori* necessary. So I shall assume that all logically impossible sentences entail a contradiction, and all logically necessary sentences are such that their negations entail a contradiction;³ and so assume that all declarative sentences which do not entail a contradiction are logically possible.

But what determines the rules of a language, and so the truth conditions of sentences, and so what entails what? Sentences of a language mean what most of its speakers (or some group of expert speakers) mean by them. Each of us learns the meanings of certain sentences by being shown many paradigm observable conditions under which those sentences are regarded as true or false, and by being told of other sentences to which a speaker is regarded as committed by uttering those sentences, and other sentences which are such that someone who utters them is regarded as committed to the former sentences. We learn the meaning of a word by being taught the difference to the meaning of a sentence made by that word playing a certain role in the sentence. By being taught the meanings of individual words and of sentences of various forms, we may then come to an understanding of the meaning of a sentence in which those words are arranged in a certain way, even if we have not been shown observable conditions under which the sentence is regarded as true or as false. Showing a language learner “observable conditions” may involve pointing to them or describing them in terms already introduced. We need to observe many different paradigm examples of observable conditions under which a sentence containing a certain word in various roles is regarded as true or false, and of the commitments speakers who use sentences containing that word in various roles are regarded as having; and this allows us to acquire an understanding of the conditions under which some new sentence containing that word would be regarded as true or false. We extrapolate, that is, from a stock of supposedly paradigm examples (of observable conditions and relations of commitment) to an

³ Robert Adams (1987, pp. 213–14) argues (in effect) that there are logical necessities whose negations do not entail a contradiction. In Swinburne (2010, pp. 318–19) I argue that the example of a sentence by which he seeks to show this does not support his view.

understanding that the sentence would be regarded as true (or false, as the case may be) under conditions sufficiently similar in certain respects to most of the paradigm examples.

This process normally leads, within limits of vagueness and minor idiosyncrasies of use, to words (and longer expressions) and sentence forms, and so sentences having a "correct" use. It leads, that is, to public agreement about what in general are the circumstances in which a given sentence would be true and the circumstances under which it would be false; and so to the commitments of sentences to other sentences. We may call a rule for what one is objectively committed to by a sentence a rule of mini-entailment. s_1 mini-entails s_2 if and only if anyone who asserts s_1 is thereby (in virtue of the rules for the correct use of language) committed to s_2 . s_1 entails s_n if they can be joined by a chain of mini-entailments, such that s_1 mini-entails some s_2 , s_2 mini-entails some s_3 , and so on until we reach a sentence which mini-entails s_n .

Given this agreement, we are then in theory in a position to determine the logical necessity, possibility or impossibility of sentences. To show some sentence s to be logically impossible we need to find an agreed chain of mini-entailments from s to a contradiction; and to show s to be logically necessary we need to find such a chain from *not-s* to a contradiction. Getting agreement that such a chain has been found is, however, often a difficult matter. An opponent of the claim that s entails a contradiction may challenge some suggested link in the chain – say the suggestion that p mini-entails q , by claiming that q is not something to which anyone is committed when using p in the correct sense. This disagreement may be overcome if the proponent of the claim that s entails a contradiction can get his opponent to recognize some r such that p mini-entails r and r mini-entails q . Or the disagreement may be bypassed if the proponent can find a different chain of mini-entailments from s to a contradiction which an opponent will recognize as such.

A sentence is logically possible if it does not entail a contradiction. Of course any logically necessary sentence is logically possible. But to show some other sentence to be logically possible (and so logically contingent) may be an even more difficult matter than to show a sentence to be logically impossible or necessary. Sometimes it is very obvious that some sentence does not entail a contradiction, and so is logically possible. A true sentence entails no contradiction, and if it is obvious that some sentence (e.g. "my desk is brown") is true, then it is obvious that it is logically possible. Sometimes too it is very obvious that some sentence, which may be false, entails no contradiction (e.g. "my desk is red"). And more generally it is sometimes very obvious that some description of a world very different from

our world entails no contradiction. To show some disputed sentence *s* to be logically possible requires showing that it is entailed by a sentence *r* which disputants agree to be logically possible without this needing to be shown by argument. The argument then consists in showing that *r* entails *s* by a sequence of agreed mini-entailments. For if *r* does not entail a contradiction, neither does any sentence entailed by *r*. For example, someone may try to show that “there are two spaces” – a space being a system of places each of which is at some distance in some direction from each other place of the system and from no other places – is logically possible, by describing in detail a situation under which it would be true.⁴ That is, they claim that the latter description entails that there are two spaces; and that since the latter description is logically possible, “there are two spaces” is also logically possible.

However, the use of these procedures to determine logical possibility presupposes that it is clear what are the truth and falsity conditions of sentences, and which sentences mini-entail other sentences. But the language-learning process, which normally produces very similar understandings of meanings in members of a language group, sometimes produces somewhat different understandings of these conditions and entailments in different sub-groups. This may occur because different learners learn meanings from somewhat different paradigm examples; and when this is recognized, language users can acknowledge that the same word or sentence has more than one meaning. But it may also occur even when both sub-groups acknowledge the same paradigm examples of observable conditions and commitments. And then it sometimes happens that one of two sub-groups objects that the sense in which some word (or longer expression) derived from the same paradigm examples (of observable conditions and mini-entailments) used by the other sub-group is not a real or legitimate sense of that word, in that its use in that sense entails contradictions. Or one of the sub-groups may object that the sense in which the word (or longer expression) is used by the other sub-group is not the sense implicit in some of the paradigm examples. It is objections of these two kinds that produce the disputes about the meaning of “personal identity.”

Most of us have been taught the meaning of the expression “is the same person as the person who” or its more natural equivalent “is the person who” by many observable paradigm examples of the same kind (e.g. “this is the same person as the person you saw last week,” “you are the person who had a headache only thirty seconds ago”) and many similar paradigm

⁴ For an argument of this form in favor of two spaces being a logical possibility, see Quinton (1962).

mini-entailments (e.g. “A is the same person as B” and “B is the same person as C” mini-entail “A is the same person as C”). We will all recognize most of these observable paradigm examples of “same persons” as examples of persons with strongly continuous bodies, memories and character. So some philosophers provide an analysis of the meaning of “ P_2 is the same person as P_1 ” in terms of P_2 having a body, memory and character strongly continuous with those of P_1 . That is, they advocate the complex theory of personal identity as a conceptual truth, in my sense a logically necessary truth. But others of us (including myself) think that it is not the normal sense of “ P_2 is the same person as P_1 ”; and that in the normal sense this expression designates a continuing identity of a different kind which normally underlies the strong physical and mental continuities but is not constituted by them and can occur without them. That is, we claim, the simple theory of personal identity is a conceptual truth, in my sense a logically necessary truth.

The only way to resolve this disagreement is by persistent continuing use of the methods described earlier. Advocates of the complex theory as a conceptual truth try to get us to recognize the logical impossibility of a personal identity independent of strong mental and physical continuities. They do this by trying to show that some sentence using the expression “is the same person as” in a different sense from theirs entails a contradiction which would not arise if the expression were used in their sense. For example, they may claim that “Socrates is the same person as the mayor of Queenborough, but has none of the same brain, memory or character as the mayor,” together with what they may claim to be a necessary truth “no one should be punished for any act which they cannot remember doing” entails “both {the mayor should be punished for any immoral acts of Socrates} and not-{the mayor should be punished for any immoral acts of Socrates}.”⁵ If they can get us to recognize this in one case, then they may get us to recognize that other sentences where the expression “is the same person as” is used in a sense other than their sense will have the same consequences, and so to see that any such sense of the expression is not a legitimate one. We who claim that the simple theory is a conceptual truth are of course likely to deny, with respect to the example just discussed, either that the first conjunct of the purported entailment (“the mayor should be punished for any immoral acts of Socrates”) is indeed entailed, or that “no one should be punished for any act which they cannot remember doing” is a necessary truth.

⁵ This example is of course a formalized version of Locke’s argument for the necessity of same memory (which he calls “same consciousness”) for personal identity (Locke 1975 [1690], II, xxvii, 19).

We advocates of the second sense are happy to acknowledge that the sense in which our opponents understand “is the same person as” – that is, as something like “has a body, memory and character strongly continuous with” – is a perfectly legitimate sense of an expression, but claim that it is not the normal sense of “is the same person as.” This is because – contrary to the claims of the advocates of the complex theory – it is not compatible with the sense implicit in some of the paradigm examples of personal identity by which the expression has been introduced into language. Some of these examples concern our opponents’ own identity. They must recognize that they themselves often have streams of overlapping experiences. For example, the second half of the experience of a pain during a “specious present” may overlap with the first half of an experience of some noise; this noise may continue for a short while (during several overlapping periods of specious present), and overlap with a certain tactual experience, and so on. It is a paradigm example of personal identity that two overlapping conscious events are experiences of the same person, from which it follows that any stream of such events are also experiences of the same person.⁶ Then they must recognize some very recent past experiences which they remember so vividly that it is obvious that they occurred (e.g. “you are the person who had a headache thirty seconds ago”); and it is obvious that – as Reid (2002 [1785], III.4) put it – “my memory testifies not only that [a certain past action] was done, but that it was done by me who now remembers it.” We thus point out to any opponent that some of the paradigm examples of personal identity that he must recognize are ones in which he has a direct awareness of personal identity; and what the awareness is of is not continuity of body, memory and character, but something which can only be described as an awareness of himself as a continuing subject of experience. Once we have focused on the paradigm examples of our opponent’s own personal identity over time, which give rise to the understanding of himself as a continuing subject of experience, we must get our opponent to recognize that, as with any experiences, what he is aware of (the continuity of his mental life) could occur without his subsequently remembering it. And so, more generally, we must get him to see that this continuity could occur without any of his criteria of personal identity being in any way satisfied. To do this we need to describe some situation in much detail by a sentence (normally consisting of a long conjunction, such as a thought experiment in which someone is described as surviving events of a kind

⁶ It was John Foster who drew attention to the phenomenon of a stream of overlapping experiences as the foundation of our understanding of personal identity. (See Foster 1979, 176: “it is in the identity of a stream that we primarily discern the identity of a subject.” See also Foster 1991, 246–50.)

described at the beginning of this chapter) which our opponents are prepared to recognize as logically possible; and then find a route of mini-entailments from it to a sentence that claims that some later person is the same person as an earlier person without our opponents' criteria being satisfied. Given that this is logically possible, our opponents' theory which claims that it is not logically impossible must be false.

My feeling about how this debate goes these days is that we are getting our message across. If it is admitted that it is physically possible, and so a fortiori logically possible, that a series of operations such as those described at the beginning of the chapter could occur and that it is logically possible that the person before the operations could subsequently have the experiences of the person after the operations, it then follows that being the same person as a previous person does not entail having the same brain or strongly continuous memory and character. Our opponents may, however, insist on some residual physical continuity, e.g. that the replacement of brain matter does not occur all at once. But someone softened up by physically possible stories of the kind described at the beginning of the chapter may then begin to acknowledge the logical possibility of a person acquiring a new body all at once without gradual replacement of parts; and so come to acknowledge that it is logically possible that a person could be the same person as a person at a later time without there being any continuity of body (including brain), memory or character between them. So, our opponent should recognize the second sense as the normal sense of personal identity.

The same arguments that will show that there is no contradiction in an unnamed person continuing to exist and have experiences under these circumstances, are unaffected by whom one supposes the person to be. So we may conclude that it is logically possible for me or any other human to survive total replacement of body, memory and character. Logical possibility is the kind of metaphysical possibility which can shown to be such a priori. But to determine whether it is metaphysically possible for me or any other human to continue to exist without any continuity of body, memory or character, we need to show that this is also a posteriori metaphysically possible. That, someone may claim, will depend on what sort of persons we humans are – that is, what is the essence of a human person – and that that is not something to be determined a priori.

A POSTERIORI METAPHYSICAL POSSIBILITY

I will begin my discussion of a posteriori modal claims in terms of what it is for a sentence to be a posteriori metaphysically necessary; the application to

a posteriori metaphysical impossibility and possibility will then become apparent. A posteriori metaphysical necessity is supposed to be a necessity as hard as logical necessity, yet discernible only a posteriori. No different type of necessity could be as hard as logical necessity, and so a posteriori necessity must be in some way reducible to logical necessity. And, as far as I can see from all the plausible examples of a posteriori metaphysically necessary sentences that have been adduced, the way “a posteriori” comes into it is that we need to make empirical enquiry to determine more adequately what is the substance or property or whatever about which the claim being made by the sentence is being made. When we have determined that, if the claim is metaphysically necessary, the necessity of the claim will be detectable a priori.

Sentences pick out the substances, properties or whatever with which they are concerned either by “rigid designators” (as defined by Kripke 1981, p. 48) – that is, expressions which (given that their meaning remains the same) always refer to the same substance, property or whatever, however different the world might be from how it is – or by non-rigid designators which may pick out different substances or whatever if the world is different. “Green,” for example, is a rigid designator because it always refers to the color green, whereas “Amanda’s favorite color” is a non-rigid designator because it would refer to a different color if Amanda had different color preferences from her present ones. I will call a rigid designator ϕ an “informative designator” if we can (when favorably positioned, faculties in working order, and not subject to illusion) recognize when something is ϕ and when it is not, merely in virtue of knowing what the word ϕ means. I would not understand the word “green” unless (when the stated conditions are satisfied) I could recognize when an object is green and when it is not. When our referring expressions are informative designators (or can be defined in terms of informative designators), we know the necessary and sufficient conditions for the things referred to to be what they are; and so, I shall say, we know the essence of what is being designated. When all the designators in a sentence are informative (or can be defined in terms of informative designators), it is a pure a priori exercise to determine whether the sentence is logically necessary or whatever. When we know what we are talking about, mere thought can show what that involves. Mere thought, for example, I suggest, can show that “all trilateral figures are triangular” (a “trilateral” figure is a closed surface bounded by three straight lines; a triangle is a closed surface bounded by straight lines and having three interior angles) or “no surface can be both green and red all over” are logically necessary; and we can now see that this is because “surface,” “straight,” “green,” “red,” etc. are all informative designators, and so

understanding the sentences involves knowing the essences of what is being referred to and so comprehending fully what the sentences are claiming.

But there are rigid designators which refer to substances or whatever, such that a speaker can understand to what they are referring on some occasions when the thing exhibits certain non-essential features, without knowing the essence of what is referred to. Clearly one will not understand what some rigid designator means (what its role is in the language) unless one knows how to use it on some occasions, and understands the kind of thing to which it is used to refer and so the kind of criteria by which to distinguish one thing of that kind from other things of that kind; but language users may not be able in practice to use these criteria to determine to which thing of that kind it is referring. But in that case one would not be able to recognize that thing on occasions other than the ones on which it exhibits those non-essential features (its “stereotype”). I shall call such designators for which the criteria of “informative designators” are not satisfied (and cannot be defined in terms of such designators), “uninformative designators.”⁷ A sentence is then a posteriori metaphysically necessary if it would be logically necessary when we substitute informative designators (or expressions definable in terms of these) for its uninformative designators.

Thus – to use the example discussed by Putnam (1975) – the word “water” as used in the eighteenth century was an “uninformative designator.” This is because although people used “water” as a designator of a stuff, and so knew that to be the same stuff something would have to have the same chemical essence, they picked out a volume of stuff as water in virtue of its superficial contingent properties (being liquid, in our rivers and seas, etc.), yet – in ignorance of what that chemical essence was – they would not be able to recognize it on occasions when it did not have those superficial contingent properties. So they were unable to say whether or not sometimes stuff found elsewhere than in our rivers and seas was water or not. When people discovered that chemical essence (H_2O), they could then recognize whether stuff not in our rivers and seas was water or not. Hence, since the claim being made about water is a claim about H_2O , we can substitute “ H_2O ” for “water” in “water is H_2O ” (as used in the eighteenth century) and the sentence then reduces to a logically necessary truth. (I assume here that “ H_2O ” can be defined in terms of informative designators: that is, in

⁷ Similar distinctions to my distinction between “informative” and “uninformative” designators are those made by Chalmers (1996) between expressions with “primary intensions” and ones with “secondary intensions,” and by Bealer (1996) between “semantically stable” and “semantically unstable” expressions.

terms of such expressions as “mass,” “volume,” “smaller by 10^{-1} than” and so on.) Somewhat similar is the sentence used by Kripke (1981, p. 100) to illustrate a posteriori metaphysical necessity, “Everest is Gaurisanker,” where – Kripke supposes – “Everest” was used by early explorers to designate a mountain having a certain shape when seen in the distance from Tibet and “Gaurisanker” was used to designate a mountain having a certain shape when seen in the distance from Nepal. The explorers understood the sentence “Everest is Gaurisanker” because they understood what it would be like for the two mountains to be the same – it would consist in their being made of the same chunk of rock. The rock of which each mountain was made constituted its essence. But they did not know whether the two mountains were the same (whether they had the same essence), and it required empirical investigation to discover that they did. Once they knew what each mountain essentially was, they knew that the claim being made by the sentence was necessarily true with a necessity as hard as that of “all squares have four sides.” Thus – in my terminology – they could know that there is an informative designator of the form “mountain made of such and such rock” which can be substituted for both “Everest” and “Gaurisanker,” so that the sentence has the form of an identity sentence “*a* is *a*” and so is logically necessary.

The application of my account of a posteriori metaphysical necessity to a posteriori metaphysical impossibility and possibility should now be evident. The metaphysical modal status of a sentence is its logical status when informative designators are substituted for uninformative ones. Even if we cannot find out what is the essence of some substance or whatever, our understanding of how to use the designator may give us enough knowledge of the kind of essence involved to enable us to see or deduce the modal status of some sentence using it. For example, even if we do not know the essence of water, we can see that “water is the number 42” is impossible. But only if all designators in a sentence are informative (or can be analyzed in terms of informative designators) is it guaranteed that mere a priori reasoning can determine its logical status.

Many of the words by which we pick out properties are informative designators (“green,” “square,” “has a length of one meter,” etc.). And many words by which we pick out properties, which are not themselves informative designators, can be analyzed in terms of them – e.g. “has a length of 10^{-18} meters” can be defined in terms of the informatively designated property “has a length of one meter” and eighteen applications of the informatively designated relation of “being shorter by one-tenth than.” However it seems that there are at present some substances which we can only pick out by

uninformative designators. For example, it is an unresolved issue (French 2006) whether some fundamental particles, such as quarks and electrons, are the particles they are merely in virtue of their properties (such as mass and charge, and causal relations to other particles), or whether they are what they are partly in virtue of the particular matter of which they are made. For this reason we do not know what are the necessary and sufficient conditions for some fundamental particles to be the ones they are (that is, we do not know their essence) and have to pick them out by uninformative designators.

Now what sort of designator is “I,” or “Richard Swinburne,” as used by me? These seem to be informative designators. If I know how to use these words, I cannot be mistaken about whether or not they apply to a certain person – given that I am favorably positioned (e.g. his body is my body), with faculties in working order, and not subject to illusion. And when I am considering applying these words to a person in virtue of his having some conscious event, these conditions will be maximally satisfied and no mistake is possible. I am, in Shoemaker’s (1994, p. 82) phrase, “immune to error though misidentification.” I cannot know how to use the word “I,” recognize that someone is having some conscious event (e.g. a pain) and still wonder whether it is I who am having that event or someone else, in the way that an early explorer could know how to use the word “Everest,” and yet wonder whether the mountain at which he is looking from Nepal is Everest. My knowledge of how to use “I,” like my knowledge of how to use “green” and “square,” means that, in the sense analyzed above, I know the essence of what I am talking about when I use the words. Hence if “I will exist tomorrow with a new brain” or “I will exist without any memory of my previous existence” are logically possible, they are also metaphysically possible. I claim therefore that the considerations which should lead people to conclude that such sentences express logical possibilities should also lead them to conclude that they express metaphysical possibilities. And since I can know this merely in virtue of knowing to what my use of the word “I” refers, other people can know the same about themselves. Each of us, we may properly conclude, can continue to exist without any continuity of brain, memory or character.

Of course I can still misremember what I did in the past, and indeed misremember how I used the word “I” in the past. But this kind of problem arises with every claim whatsoever about the past. “Green” is an informative designator of a property, but I may still misremember which things were green and even what I meant by “green” in the past. The difference between informative and uninformative designators is that (when my faculties are in working order, I am favorably positioned and not subject to illusion) I can

recognize which objects are correctly picked out at the present time by informative designators, but not necessarily when they are picked out by uninformative designators (in the absence of further information). And so I know what a claim made now about the past or future amounts to when it is made by informative designators (but not when it is made by uninformative designators) whether or not I have any reason for supposing it to be true. For to claim that some informatively designated object a will exist or have an informatively designated property ϕ tomorrow is just to claim that something which I can understand (a existing or being ϕ today) will hold tomorrow. It follows that I can understand what it is for me to exist tomorrow or yesterday and to have such and such experiences. Not so with Everest or water, when these things can be picked out only by uninformative designators. I do not know what would constitute a past or future substance being water or Everest if I am merely in the position of the “water” user in the eighteenth century, or the early explorers using “Everest” in the way described.

I conclude that given that each of us can come to see that it is logically possible that they can survive without any continuity of brain, body or character, in the crucial sense in which they subsequently have the experiences of the surviving persons, and so come to see that this is metaphysically possible and so come to see that the simple theory of personal identity is true.

THE HUMAN SOUL

I stated the simple view as the view that personal identity is not constituted by continuities of mental or physical properties or of the physical stuff (that is, matter) of which persons are made, but is a separate feature of the world from any of the former. But this leaves open the possibility that personal identity might be constituted by a non-physical part, a “soul.” Substance dualism holds that each human on earth consists of two parts, a body and a soul – the soul being the essential part, and the body a contingent part. On this theory, while any physical stuff of which the body is composed, and any particular physical and mental properties, are not necessary for the continued existence of a person, the continued existence of his soul is necessary. I suggest that this further step is forced upon us if we admit the logical possibility of a certain thought experiment, and the high plausibility of a certain principle about what it is for any substance to continue to exist.

The thought experiment is this. Suppose that there exists instead of our actual world W_1 , a world W_2 which is exactly the same as W_1 except in that instead of a certain person S_1 who lives a certain life in W_1 there is a person S_2 who has the same body and the same mental and physical properties

throughout his life as S_1 , but is not S_1 . And surely our world could be different in the sole respect that the person who lived my life was not me. For it is not entailed by the full description of the world in its physical aspects and in respect of which mental properties are associated with which bodies that the person who has any particular body and mental and physical properties should be me. We can see this if we imagine that before this world exists we are shown a film of what is going to happen in it; and that the film has some device for showing us what will be the mental lives of the people in the world. Each of us would still not know whether we were going to live one of the lives in this world. And if so, which one. So because W_2 can be seen, I suggest, when we reflect on it, to be logically possible; and because – as before – persons can be picked out by themselves by the informative designator “I,” W_2 is a metaphysically possible world.

But a substance at a particular time is the substance it is in virtue of its parts, what they are made of and their properties, and the relations the parts have to each other. For example, if there is a substance composed of certain fundamental particles with certain properties (including certain relations to past particles) related to each other by certain causal and spatio-temporal relations, there could not be instead of it a different substance composed of all the same particles with the same properties and relations to each other. Andre Gallois (1998, p. 251) has called the view that there could be another such substance “strong haecceitism.” He writes:

Strong haecceitism seems to me incredible. Consider a car on a parking lot. It is not at all incredible to suppose that a qualitative duplicative of the car in question might have existed even if there is no qualitative difference at any place or time as a result. It is incredible to suppose that throughout history all of the atoms that actually exist might have been configured at each time in exactly the way they are actually configured without the car on the parking lot existing.

I suggest that it follows from our very understanding of what a substance is that what Gallois describes as “incredible” is false; and in particular it follows from our understanding of what a person is that two persons could not be different if all their parts and all their properties were the same. (This does not commit me to the identity of indiscernibles, which holds that two substances are the same if they have all the same properties (in the sense of universals). It allows that two substances may be different even if they have all the same properties, so long as their parts are different.)

It follows that in the thought experiment described above, S_1 and S_2 must have different parts; and since all their physical parts are the same, the difference must arise from each having different non-physical parts, that is

different souls. My earlier thought experiments suggested that a person can continue to exist with a different body and different mental properties. The thought experiment just described in this section suggests that a person needs a particular soul in order to continue to exist. It is, however, compatible with substance dualism to hold that to exist a person needs a body, but not any particular body – though I believe that further thought experiments show that a person can exist without a body and so without any physical properties at all. So the only essential properties necessary for a person to exist are the essential properties of any soul, which – I suggest – are simply the one property of having (in some sense) a capacity to be conscious.⁸

⁸ In my book *The Evolution of the Soul* (Swinburne 1997, pp. 153–4, and pp. 327–32), I followed St. Bonaventure in analyzing a soul as a form (a collection of essential properties) instantiated in some mental stuff, soul-stuff. But it now seems to me that any stuff must be capable of being divided into smaller chunks of the same stuff; and given my view that humans (and so their souls) cannot be divided, the soul cannot be made of any stuff. It is an “immaterial particular.”