11

Personal Identity: What's the Problem? (1995*u*)

I

We are gathered here in San Marino to pay homage to the world's greatest living philosopher.[1] Saul Kripke has been aptly described as the one true genius of contemporary philosophy. He is indeed a phenomenon, nothing less, and the discipline is much the better for his contribution (only a faction of which is represented by his published work). My own intellectual development has benefitted immeasurably from my association with Kripke, first as a student, later as a colleague, always as a friend. Another great living philosopher, Woody Allen, once said that he did not wish to achieve immortality through his work; he wished to achieve it through not dying. Like Allen, Kripke will live on through his work long after most of the rest of us are forgotten.

In Woody Allen's semi-autobiographical movie, *Stardust Memories*, his character says the following:

> I've never been able to fall in love. I've never been able to find the perfect woman. There's always something wrong. And then I met Doris. A wonderful woman, great personality. But for some reason, I'm just not turned on sexually by her. Don't ask me why. And then I met Rita. An animal, nasty, mean, trouble. And I love going to bed with her. Though afterward I always wished that I was back with Doris. And then I thought to myself, 'If only I could put Doris's brain in Rita's body. Wouldn't that be wonderful?' And I thought, 'Why not? What the hell, I'm a surgeon.' . . . So I performed the operation, and everything went perfectly. I switched their personalities. . . . I made Rita into a warm, wonderful, charming, sexy, sweet, giving, mature woman. And then I fell in love with Doris.

This fictitious tragedy raises a host of philosophical issues. The central issues concern the irrational nature of human sexual attraction and romantic love, and the often troubling relationship between the two. The dialogue also raises moral issues about the treatment of people as means rather than as ends in themselves, the

[1] The present chapter was delivered (in part) at the University of San Marino International Center for Semiotic and Cognitive Studies Conference on Saul Kripke's Contribution to Philosophy, May 1996, and incorporates portions of my 'Trans-World Identification and Stipulation,' *Philosophical Studies* (forthcoming). I am grateful to Anthony Brueckner and Jill Yeomans for their comments. The chapter is dedicated to a remarkable woman, Sandy Shaffer, who has survived her challenges with tenacity and inspiring grace.

objectification and victimization of women, and related issues. The passage also concerns the traditional philosophical problem of the identity of a person through change. With profound apologies to the reader, the present chapter is concerned exclusively with personal identity. I shall argue that the traditional philosophical problem dissolves. Recent discussion has tended to focus on the question of 'what matters' in survival, with less attention paid to the original question of what makes someone the very same person even through change. This may because it is widely believed that strict survival—genuine personal *identity*—is not what is fundamentally important and not what ought to concern us. Though I remain doubtful that this has been successfully argued, I shall not discuss the issue here. If I am correct, there is a better reason for dismissing the question of what personal identity consists in.

Others before me have rejected the problem of personal identity (or more generally, the problem of the identity of a thing through change) as a pseudo-problem, on the ground that it presupposes the questionable doctrine that a person is constituted by *stages* (*phases, temporal parts*), which are supposed to be portions of that person's life history.[2] Once this doctrine is rejected, it is argued, it follows immediately that there is no genuine problem about formulating principles of unification that specify which series of person stages constitute genuine persons, as opposed to gerrymandered non-persons. My objection to the alleged problem of personal identity has virtually nothing to do with this one, which seems to me to be wide of the target. I have no quarrel to make against stages or phases. No doubt much of what has been supposed about them is simply wrong, but that is not sufficient reason to doubt their existence.[3] More important, the typical puzzle cases for personal identity can easily be set out without any appeal, explicit or implicit,

[2] For an elegant presentation of the problem of personal identity by means of person stages, see John Perry's introduction, 'The Problem of Personal Identity,' to his valuable edited collection, *Personal Identity* (Berkeley: University of California, 1975), pp. 3–30.

[3] Those who frame the problem of personal identity in terms of person stages tend towards the view that stages are conceptually prior to, or metaphysically more fundamental or real than, the continuants that they constitute through time. They often base their view on Leibniz's Law: *If x is the same thing as y, then x is exactly like y in all respects*. Being a law, this holds for any time *t*. The stage theorist presupposes an alternative, incorrect temporal generalization: For any *pair* of times *t* and *t′*, if *x* is at *t* the same thing that *y* is at *t′*, then *x* is at *t* exactly like *y* is at *t′*. Alternatively (or in addition), some stage theorists misunderstand what it is to have a property at a time *t*. The stage theorist presupposes that to be such-and-such at *t*, for any time *t*, entails being such-and-such *simpliciter* (whereas, in fact, to-be such-and-such *simpliciter* is to be such-and-such at the present time). The erroneous temporal over-generalization of Leibniz's Law, and equally the misunderstanding of what it is to have a property at a time, exclude the possibility of genuine change in an enduring object. Each thus raises a pseudo-issue of how, or in what sense, a single thing can be such-and-such at *t* and not be such-and-such at *t′*. The stage theorist's answer is that only part of the thing is such-and-such, while another part is not such-and-such. *Cf.* David Lewis, *On the Plurality of Worlds* (Oxford: Basil Blackwell, 1986), at pp. 202–204; and Mark Johnston and Graeme Forbes, 'Is There a Problem about Persistence?' *Proceedings of the Aristotelian Society*, supplementary v. 61 (1987), pp. 107–155. Forbes defends an account that I favor of what it is to have a property at a particular time (pp. 140–142). *Cf.* my *Frege's Puzzle* (Atascadero, Ca.: Ridgeview, 1986, 1991), at pp. 24–43; and 'Tense and Singular Propositions,' in J. Almog, J. Perry, and H. Wettstein, eds., *Themes from Kaplan* (Oxford University Press, 1989), pp. 331–392. While opposed to Lewis's postulation of stages, Johnston joins Lewis in objecting to the account I take, on the ground that

to the notion of a person stage (or anything similar). Something different must be said—or at least something more—if the problem, so formulated, is to be rejected as illegitimate.

The aspect of the problem that I discuss here is connected to a couple of doctrines recently brought into prominence by Saul Kripke's influential monograph, *Naming and Necessity* (Cambridge, Mass.: Harvard University Press, 1972, 1980). First is the doctrine of *individual essentialism*, according to which some properties of individuals are such that those individuals could not exist without those properties. To put it another way, there are properties that certain individuals have in every possible world in which those individuals exist. Second is Kripke's claim that possible worlds are not discovered like planets but 'stipulated.' In previous work, I have defended the idea that in whatever sense it is correct and useful to recognize possible worlds as entities, it is equally correct and useful to acknowledge that there are also *impossible worlds*.[4] My doctrine of impossible worlds has proved controversial, at least partly because it has seemed unclear whether such an apparatus has any philosophical utility. I here apply the doctrine, in a manner that I hope will prove its mettle, to the traditional problem of personal identity. I shall also bring the controversy of Haecceitism vs. Anti-Haecceitism, and the distinction between reducibility and supervenience, to bear on the problem.

To see how the alleged problem of personal identity be presented without appealing to person stages and principles of unification, one need only look to the nonphilosopher who does not know from person stages. Woody Allen's character tells us that he has performed a complex surgical procedure on Doris and Rita,

any time, past or future, is as real as the present. I respond that the past *was* real but is so no longer, and the future *will be* real but is not so yet. The present is currently real in a way that the past and the future are not. This truism is unaffected by the context-relativity of the words 'now', 'past', 'present', etc. *Cf.* my 'Existence,' in J. Tomberlin, ed., *Philosophical Perspectives*, 1 (Atascadero, Ca.: Ridgeview, 1987), pp. 49–108, especially at 73–90.

Johnston defends an account according to which having a property at t is having the property in a certain manner (being such-and-such 'in the t-mode,' as it were). Though this is virtually derivable as a special case from the account that Johnston joins Lewis in rejecting, Johnston instead takes his account to be superior in allegedly according the past and the future the same ontological status as the present. Johnston's account has the significant disadvantage that it applies only to temporal qualifications of subject-predicate sentences, e.g. 'In 1987, *a* was such-and-such', and does not directly provide an interpretation for sentences like 'It will rain tomorrow', 'In 1987, there was something such that . . . it . . .', etc. For those sentences to which his account applies, Johnston ultimately falls back on familiar tense-logical semantics (p. 128). The latter holds that to be such-and-such *simpliciter* is to be such-and-such at the present time, and more generally, that truth *simpliciter* (i.e. in reality) is truth at the present time, whereas Johnston evidently means to reject the very idea of being such-and-such *simpliciter*. Why then not also reject the idea of reality, and replace it with different ways of being real (truth in the t-mode, truth in the t'-mode, etc.)?

[4] 'How *Not* to Derive Essentialism from the Theory of Reference,' *Journal of Philosophy*, 76, 12 (December 1979), pp. 703–725, at 723–724n; *Reference and Essence* (Princeton University Press, 1981), section 28 (especially pp. 238–240); 'Impossible Worlds,' *Analysis*, 44, 3 (June 1984), pp. 114–117; 'Modal Paradox: Parts and Counterparts, Points and Counterpoints,' in P. French, T. Uehling, and H. Wettstein, eds., *Midwest Studies in Philosophy XI: Studies in Essentialism* (Minneapolis: University of Minnesota Press, 1986), pp. 75–120; 'The Logic of What Might Have Been,' *The Philosophical Review*, 98, 1 (January 1989), pp. 3–34; 'This Side of Paradox,' *Philosophical Topics*, 21, 2 (Spring 1993), pp. 187–197.

interchanging the brains between their two bodies, and consequently interchanging also what I shall call their 'psychologies'—that is to say, their personality and character traits, their beliefs, attitudes, wishes, hopes, fears, memories, abilities, talents, habits, mannerisms, and the like. The standard philosophical question raised by the incident involving Doris and Rita—*the d/r Incident*, as I shall call it—is sometimes framed in terms of how one of the two person stages at some time t immediately after the surgery should be related to various person stages prior to the surgery in order for the stages to qualify as stages of a single person. But the question may be framed instead in terms of the identities of the two women to emerge from the surgery. Consider the woman with whom Allen has now fallen in love—she who now occupies what used to be Doris's body but who now has what used to be Rita's brain. Is that woman Doris? Is she Rita? Or is she perhaps someone else—call her 'Dorita'—who was created in the process, while Doris and Rita were destroyed?

At least three philosophical questions must be distinguished here. The issue of whether the woman in question is Doris or Rita, or neither, is *the primary question* about the D/R Incident.[5] In addition there is the question of how the correct answer to the primary question is determined. This meta-question is often put by asking for (and very often by demanding) a *criterion*, or *criteria*, that settle the primary question. The question bifurcates into two separate questions, which, although they may call for distinct answers, have often been blurred together. First, there is *the epistemological question* of how, or by what means or evidence, one is supposed to come to know or to discover the answer to the primary question about the D/R Incident. Second, and more fundamental, is *the metaphysical question* concerning the correct answer to the primary question, of what makes it the correct answer. In virtue of what fact or facts is it, and not its rivals, the right answer to the primary question? In short, what is it to be the same person? Although each of the three questions has been posed as 'the problem of personal identity,' it is the metaphysical question that has the strongest claim to being *the* problem of personal identity, as the phrase is traditionally meant.[6]

Although the demand for a criterion of personal identity is frequently made, the relevant notion of an identity criterion is usually not made precise. One way of understanding what a personal-identity criterion is that seems to fit much of the literature takes it to be a *trans-temporal link* that connects a person from one time to a person of another and thereby determines that they are the same. More precisely, on this interpretation a *criterion for personal identity* is an ordered triple consisting of a sortal property F and a pair of binary relations R and R', other than personal identity itself, such that it is necessary that for any persons x and y and any times t and t' such that x exists at t and y exists at t', x is the same person at t that y is at t' if there is some F (i.e. something of sort F) to which x bears R at t and to which y bears R' at t'. In most cases, but not all, the intent is better captured by strengthening the 'if' to 'if and only if'. Either way, the particular F is supposed to serve as the link (*via* the relations R and

[5] The classical discussion of this question is Sydney S. Shoemaker, *Self-Knowledge and Self-Identity* (Ithaca: Cornell University Press, 1963), pp. 23f. See also Shoemaker's 'Personal Identity and Memory,' *Journal of Philosophy*, 56 (1959), pp. 868–882; reprinted in Perry, ed., *Personal Identity*.

[6] *Cf.* Shoemaker, *Self-Knowledge and Self-Identity*, pp. 2–3ff.

R') that determines personal identity.[7] A memory-based criterion results by letting F be the sortal *experience token* letting R be the relation of *remembering*, and letting R' be the relation of *experiencing*. According to this criterion, by necessity, x is the same person as an earlier person y if (and only if) x remembers having some experience token of y's. Here the remembered experience links x to y across time.[8] A body-based criterion, by contrast, results by letting F be the sortal *body*, and both R and R' be the relation of *being the functional owner of*—in this case, the relation *u is the person whose body is v*. According to this criterion, by necessity, x and y are the same person if they are linked by having the same body across time. (The reader is invited to verify whether other criteria that have been proposed can also be put into the same general form involving the existence of a trans-temporal link.)

The ambiguity in the meta-question may be traced to a choice regarding the kind of necessity involved in the notion of a personal-identity criterion. The epistemological meta-question results by taking the necessity to be epistemic. The metaphysical meta-question results by taking the necessity to be alethic rather than epistemic. In the former case, the trans-temporal link is the epistemic basis for the judgment of personal identity over time. In the latter case, the link is the *metaphysical* basis for the *fact* of personal identity. Personal identity would thus *consist in* the existence of an appropriate trans-temporal link.

II

Allen says that he made Rita into the ideal mate he was seeking, and so, naturally, he has fallen in love with Doris. By putting things this way, he is evidently presupposing the body-based criterion for personal identity, according to which the woman who now has what was previously Doris's body is Doris, and the woman who now has what was previously Rita's body is Rita. If Allen had instead presupposed a psychology-based criterion—such as the memory-based or a personality-based criterion—he should have described the outcome of the D/R Incident by saying that he has made *Doris* into an ideal mate, but (alas) has fallen in love with *Rita*. Allen

[7] The resulting condition for personal identity is the *relative product* of R and the converse of R'. Although the relation of personal identity between x and y is here taken to be a trans-temporal relation, holding between objects across times (more accurately, holding among a quadruple of a person x, a time t, a person y, and a time t'), each of the criterial relations R and R' obtains between objects at a single time. For discussion of an analogous account of cross-world relations, see my *Reference and Essence*, section 13, pp. 116–135. With some ingenuity, other sorts of identity criteria, even criteria for identity at a time (as opposed to identity across time), might also be put into the same general form. For example, the traditional criterion for the identity of sets may be put: $x = y$ iff there is a particular membership m such that x has m and y has m.

[8] This memory-based criterion is not a counter-instance to the observation made in the preceding note that each of the criterial relations R and R' obtains between objects at a single time. The remembering of the experience takes place at a single time when the experience is already past. Although the remembered experience is no longer current, and hence in some sense no longer 'real,' the person remembering it enters into a relation with it while remembering it, precisely by remembering it. (Alternatively, one might let F be the sortal *biographical event,* R' be the relation of *being the principal figure involved in* a particular event, and R be the relation of *remembering being the principal figure involved in*.)

puts things as he does not because he is a closet materialist, but because he is a brilliant humorist. For some reason, putting things the other way spoils at least some of the humor of the monologue. This may reflect a natural tendency to identify people by their bodies. This tendency may obtain among most people, evidently including even the cleverest and most philosophical of non-philosophers.

In order not to beg the primary question in setting out the philosophical conundrum, philosophers have invented an artificial terminology better suited to philosophical debate. Philosophers call the person who now has what used to be Rita's brain in what used to be Doris's body 'the Doris-body-person', and we call the person who now has what used to be Doris's brain in what used to be Rita's body 'the Rita-body-person'. We may then pose the question: Is the Rita-body-person Rita, or is she Doris? Allen presupposes that the Rita-body-person is still Rita and the Doris-body-person is still Doris. The artificial terminology allows for a way of putting things that neither presupposes nor excludes any criterion of personal identity. We may say, neutrally, that the Rita-body-person is now an ideal mate, but Allen has fallen in love with the Doris-body-person. The primary question may be posed by asking whether Allen has fallen in love with Doris or Rita. That's not comedy; it's philosophy. The joke has been butchered, but the conundrum has been given life.

The different ways of making the identifications are conceptually at odds. They carry with them different conceptions of the changes that have taken place in Allen's victims. On the psychology-based identifications, Doris and Rita retain their brains intact, and therefore also their psychologies. They have *exchanged bodies*. More accurately, their bodies have been interchanged by Allen. Body swapping would no doubt require a variety of adjustments in one's life, some quite radical. Other than the ensuing psychological adjustments, however, on the psychology-based identifications Doris and Rita remain fundamentally unaltered psychologically. This way of making the identifications is committed to making sense of the alleged phenomenon of re-embodied minds or spirits—or to put it perhaps less tendentiously, of re-embodied persons. (It does not require the possibility of altogether disembodied persons, let alone of persons without brains.) By contrast, on the body-based identification, Doris and Rita retain their bodies while having exchanged brains. Each of their individual psychologies has thereby undergone a radical transformation. Although the two women have the same bodies, they are not at all the same as they used to be. One might even say (as Allen does) that the women have traded personalities. Doris now has the personality that was previously Rita's while Rita now has the personality that was previously Doris's. As persons, they have been psychologically *altered* or *modified*. Rita has been transformed into an ideal mate, and Doris has been modified to such an extent that Allen is now obsessed with thoughts of her. This is a very different interpretation or conceptualization of the changes in Doris and Rita. The psychology-based identifications carry with them the ideology of relocation, Allen's body-based identifications the ideology of transmutation. And, of course, the Dorita hypothesis carries with it the ideology of annihilation.

The two different ways of making the identifications and are not merely alternative descriptions differing in conceptual flavor but otherwise equally acceptable. The two conceptualizations are logically incompatible. In effect, they present entirely

different scenarios. At least one of them is mistaken. One is a misdescription of the situation. This is proved by the transitivity of identity. On the body-based identification Doris = the Doris-body-person, whereas on the psychology-based identifications Doris = the Rita-body-person. Yet it is clear that the Doris-body-person ≠ the Rita-body-person. Therefore at least one of the criteria gets things wrong. Or again, on the body-based identifications the victims retain their bodies while exchanging psychologies, whereas on the psychology-based criteria the victims retain their psychologies while exchanging bodies. Since it is logically impossible to retain one's body (or one's psychology) while also trading it for another, of necessity one or the other of these accounts of the D/R Incident is incorrect. Whichever description is correct (if either is), there is indeed an alternative but equally correct description. For whether it is correct to say of Doris and Rita that they have retained their bodies while switching their psychologies or *vice versa*, it is equally correct to say that the Doris-body-person has what was previously Doris's body and what was previously Rita's psychology whereas the Rita-body-person has what was previously Rita's body and what was previously Doris's psychology. This is the philosophically neutral way of describing the D/R Incident. It is neutral because it is incomplete. It fails to state all the relevant facts. In particular, it does not identify either the Doris-body-person or the Rita-body-person with either Doris or Rita. By design it leaves the identities of the Doris-body-person and the Rita-body-person wide open. To identify is to risk error.

The incompatibility between the two ways of making the identifications will perhaps strike the reader as trivial. That is for the good. The point *is* trivial. But it is often obscured in discussions on the topic—and that is reason enough for me to emphasize it here. It is extremely important to be clear on this point if we are to make any progress toward solving the problem of personal identity.[9]

Allen's joke exploits the body-based identifications. We, however, are not writing comedy; we are doing extremely serious philosophy. And fortunately, though not always easy to do, philosophy is always a good deal easier to do than comedy. Philosophically, the psychology-based identifications seem considerably more plausible than the body-based identifications—not as funny, but more plausible.

[9] Derek Parfit claims, in *Reasons and Persons* (Oxford University Press, 1986), at pp. 242–243, 259–260, that the different ways of making the identifications in puzzle cases of personal identity are 'merely different descriptions of the same outcome,' while explicitly denying that the competing descriptions are incompatible. His argument evidently assumes that if facts of one kind (e.g., personal identity) are reducible to, and hence not 'further facts' beyond, those of another kind (psychological and/or bodily continuities), then the former are somehow illusory or unreal—or at least less real—so that the latter are compatible with utterly different ways of fixing the former. I disagree. If facts of one kind reduce to facts of another, then the latter determine the former. And if one sort is real, then so is the other.

Bernard Williams, in 'The Self and the Future,' *The Philosophical Review*, 79, 2 (April 1970), also in his *Problems of the Self* (Cambridge University Press, 1973), pp. 46–63, presents a rich account (pp. 52–55 in *Problems of the Self*) of the conceptual distinctions between the two different ways of making the identifications in a case like the D/R Incident. He notes that a description of the incident in completely neutral terms seems to lead naturally to the psychology-based identifications, but he also says of the situation given by the body-based description that it is in fact the same incident 'differently presented.' Unlike Parfit, Williams explicitly adds that the two 'presentations' thus lead to contrary conclusions (p. 61). Indeed, Williams sees the incompatibility of these otherwise plausible presentations of the incident as producing a philosophical quandary.

This is not to say that some psychology-based criterion is correct in general. Even if it is taken as settled that the body-based identifications are clearly incorrect, it is arguable that the person who now has what had been Rita's brain in what had been Doris's body is neither Doris nor Rita but Dorita. Even if the psychology-based identifications are not decidedly vindicated in the D/R Incident, the body-based identifications seem decidedly refuted. Moreover, the psychology-based identifications do not seem at all implausible. If the D/R Incident presented all there were to the problem of personal identity, we might as well move on to discuss the more intriguing issues raised by the D/R Incident. But it does not.

III

Suppose that instead of transplanting brains, Allen had made use of the BW device. Although it is sometimes referred to as 'the brain washer,' the initials 'BW' actually refer to the device's inventor, Bernard Williams. As Williams describes the device, it extracts 'information' from a person's brain—or, as we might put it nowadays, it extracts the operating system, the memory, and all the stored data and software. Exploiting the latest in digital technology, the device stores that information while the brain is repaired. Once the brain is repaired, the device is set it in the reverse mode, whereby it copies the information back into the brain, restoring the brain to exactly the same state it was in when the information was extracted.[10] The BW device is especially useful when removing a brain tumor that is located perilously close to brain areas intricately tied to certain higher cognitive phenomena (including certain abilities, long-term memories, vocabulary, and capacity for speech, sense of humor, and various other aspects of a personality). On one or two occasions, the BW device successfully extracted information from a dying brain and replaced it in an artificial brain that had been surgically implanted in place of the old one. The BW device also has the capability simply to render the brain a *tabula rasa*. If the information had been correctly extracted and stored, the washed brain can be restored to its former state. Although the prospect has been condemned as unethical by extremists, it is theoretically possible using two BW devices simultaneously to interchange all of the information of two brains.

Suppose Allen had done exactly that to Doris and Rita. Let us call the original D/R Incident 'D/R-1' and this new scenario 'D/R-2'. We may pose our three questions with regard to D/R-2: Which way of making the identifications, if either, is right about D/R-2? How is one supposed to settle the primary question? Finally, whichever way of making the identifications is correct, by virtue of what facts is it, rather than the alternative way, the right way?

D/R-2 seems to make our problem of personal identity less tractable than it first seemed. For now the body-based identifications do not seem as implausible. Interestingly, they may even seem more plausible in this case than the psychology-based identifications. Anyone who does not find them so is urged to reread Williams'

[10] *The Problems of the Self*, p. 47. Parfit's 'Branch-Line Case of the Teletransporter,' described in his *Reasons and Persons*, at pp. 199–201, is a variant of the BW device.

discussion, in which he deftly uses a puzzle case like D/R-2 to argue that one cannot legitimately dismiss the body-based criterion as cavalierly as one might be inclined to do.[11] Intuitive support for the body-based identifications is provided by supposing that one's own brain were drastically altered through a BW device, and considering how one views the further prospect of the resulting person's being painfully tortured. Perhaps Allen's body-based identifications are the right ones after all. The two meta-questions seem more pressing.

The very fact that our intuitions may diverge between D/R-1 and D/R-2 is itself an extremely important aspect of the problem. Presented in the right light, D/R-1 and D/R-2 bring our intuitions into direct conflict, thereby creating an especially perplexing conceptual difficulty. The tension between the intuitions that are operative in D/R-1 and D/R-2 shows that the problem of personal identity is not so easily laid to rest.

If D/R-1 stacks the deck in favor of psychology-based criteria and D/R-2 stacks the deck in favor of the body-based criterion, we can make our problem even more intractable by considering a case that does not stack the deck at all. In D/R-3, something mysterious happened to Doris and Rita while they slept, with the result that the Doris-body-person awoke with what was previously Rita's psychology and the Rita-body-person awoke with what was previously Doris's psychology. Allen did not interchange their brains. He did not apply BW devices to exchange information between their brains. He did not do anything to them. Someone else—or something else—did. Perhaps it was the fruition of a curse against their ancestors in ancient Egypt. Perhaps space creatures zapped them with alien rays. Perhaps it was the magic fulfillment of a mutual wish to trade places. Never mind what it was. Allen has fallen in love with the Doris-body-person. But who is that?

Consider now the primary question and the two meta-questions concerning D/R-3. Our intuitions seem to offer decidedly less assistance in this case than they did before.

[11] Williams argues that a neutral 'presentation' of cases like D/R-1 and D/R-2 leads to the psychology-based identifications, whereas a specially designed alternative presentation leads to the opposite identifications. He evidently concludes that the case for psychology-based identifications is deeply inconclusive. (See note 9 above.) In noting the conceptual differences between the alternative presentations, Williams emphasizes two aspects that are prominent in his own presentations: First, in presenting the scenario in neutral terms the victims are referred to using the third-person, whereas the body-based scenario is presented as addressed to one of the victims using the second-person pronoun (and as understood by the victim using the first-person); and second, in presenting the body-based scenario, Williams makes little mention of the other victim. Williams creates the impression that these differences are crucial to the philosophical issues. These differences, however, are largely stylistic, reflecting different perspectives that Williams chose, perhaps at least in part, for dramatic effect. He could have provided a body-based presentation using the third-person perspective, or a neutral presentation using the second-person (and even the first-person, as by 'the my-body-person'), and still raise the principal philosophical questions on that basis. A more significant difference is given by the very fact that the body-based presentation explicitly includes particular identifications. By contrast with Williams, I believe that the proper lesson of his investigation is that the psychological evidence in favor of the psychology-based identifications—which is the focus in Williams's neutral description—has no force, since the very same psychological reactions would arise in Doris and Rita even if the body-based identifications prevailed. (Compare Kripke's 'schmidentity' argument strategy, in *Naming and Necessity*, at pp. 107–108, elaborated on in Kripke's 'Speaker's Reference and Semantic Reference,' in P. French, T. Uehling, and H. Wettstein, eds., *Contemporary Perspectives in the Philosophy of Language* (Minneapolis: University of Minnesota Press, 1977), pp. 6–27, especially at 16–18.)

Whatever conviction one may have had when considering D/R-1 that Doris and Rita have exchanged bodies is considerably weakened. The opposite intuitions tapped by D/R-2 now seem equally legitimate. And conversely, whatever conviction one may have had when considering D/R-2 that Doris and Rita have retained their bodies while becoming psychologically altered is also considerably weakened, in light of the equal legitimacy of the intuitions tapped in D/R-1. Concerning D/R-3, both intuitions seem equally legitimate, or equally illegitimate. There seems to be little to recommend the psychology-based criteria over the body-based criterion, or vice versa. The metaphysical metaquestion concerning D/R-3 in some sense represents the traditional problem of personal identity in its purest and least tractable form.

Some philosophers maintain that there is no determinate, objective fact of the matter (independently of any decision we may make about the case) as to whether the Doris-body-person is identical with Doris or Rita, or neither. This position, however, is not a viable option. Let us name the Doris-body-person 'Doris-bod'. There is a fact of the matter concerning whether Doris-bod is identical with Doris-bod. The fact that Doris-bod is Doris-bod is an instance of a law of logic. If there is no objective fact of the matter as to whether Doris is Doris-bod, then that yields one respect (at least) in which Doris differs from Doris-bod. For on this hypothesis, Doris-bod has the feature that there is a fact as to whether she is Doris-bod while Doris lacks this feature. But if Doris and Doris-bod are not exactly alike in every respect—if they differ in any respect whatsoever—it follows by Leibniz's Law that they are distinct persons. (Or if one prefers, it follows by the contrapositive of Leibniz's Law—see note 3 above.) And if they are distinct, then there is a determinate, objective fact of the matter after all as to whether they are identical. The same argument may be made concerning Rita and Doris-bod. As desperate as the Dorita hypothesis seems, one may be inclined at this point to run with it.[12]

[12] Parfit takes the position that there is no determinate, objective fact of the matter in some of the puzzle cases of personal identity. See for example his 'Personal Identity,' *The Philosophical Review*, 80 (January 1971), pp. 3–27; and *Reasons and Persons*, at pp. 236–243. The motion is seconded by Johnston, in 'Fission and the Facts,' *Philosophical Perspectives, 3: Philosophy of Mind and Action Theory* (Atascadero, Ca.: Ridgeview, 1989), pp. 369–397, throughout and especially at 371–373, 393; and again in 'Reasons and Reductionism,' *The Philosophical Review*, 101, 3 (July 1992), pp. 589–618, at 603. (Curiously, Parfit also says that in puzzle cases of personal identity, different ways of making the identifications are 'different descriptions of the same outcome,' and furthermore that for reasons of symmetry, the best description of the standard fission case has it that the original person is distinct from each of the two subsequent people. Each of these claims seems incompatible with Parfit's doctrine of indeterminate identity, as well as with each other. See note 9 above.) I urged a version of the proof just given against Parfit in *Reference and Essence*, at pp. 242–246 (see especially p. 242n). Philosophers who embrace, or otherwise defend, the logical possibility of indeterminate identity have gone to extreme lengths to ward off the counter-proof. Typically, they have responded by accepting that the objects in question (in our case, Doris and Doris-bod) differ from each other in the respect cited while rejecting the Leibniz's-Law inference from '*a* and *b* are not exactly alike' to '*a* and *b* are not the same thing', on the ground that the conclusion may lack truth value even when the premise is true. In his *Reasons and Persons*, Parfit endorses such a response (pp. 240–241). The response, however, requires a fundamentally counter-intuitive departure from classical reasoning. For it should be agreed that, of necessity, any *one* thing has every property it has, without exception. It follows by classical reasoning that if Doris lacks some property that Doris-bod has, then they cannot be *one* person. But if they are not one person, then they are two. (They are certainly not one and one-half persons, for example. *Cf.* my 'Wholes, Parts,

IV

D/R-1, D/R-2, and D/R-3 are distinct possibilities. Technically, though, they are not genuine, full-fledged possible worlds. Possible worlds are fully specific with respect to all questions of fact, down to the finest of details. There are numerous alternative

and Numbers,' in J. Tomberlin, ed., *Philosophical Perspectives*, 11, Atascadero, Ca.: Ridgeview, forthcoming 1997.) [*Homework exercise:* Formalize and derive the preceding argument. What inference rules and/or logical axioms are involved in the derivation? Notice also my use of the plural form '*objects* in question' and of the phrase 'differ from each *other*' in stating the typical response to the original proof. Is this usage consistent with the position stated thereby? If not, is there a coherent way to state the position, in its full generality?]

Parfit says furthermore that even if the proof that there is always a fact of the matter is correct, it only shows that in those cases in which there is no fact of the matter, it is incumbent upon us, if we wish to avoid incoherence, to create a fact by making a decision about the case at hand. This betrays a serious misunderstanding of the proof—and indeed, I believe, a fundamental confusion concerning such things as facts, decisions, and incoherence. The proof demonstrates that there is *already* a fact of the matter, quite independently of any decisions one may wish to make. In addition, a slight variation of the argument shows that it is quite impossible to make a pair of things identical (or distinct) by decision. Doris and Doris-bod are already what they are, and no decision on anyone's part can possibly affect their status with regard to the question of identity.

Johnston argues instead that even if the notion of personal identity (the notion of *same person*) is taken to be strict numerical identity restricted to persons, and even if strict identity is determinate for every pair of objects, there are nevertheless cases in which it is indeterminate whether a is the same person as b owing to an ambiguity in the word 'person'. His position appears to be that there are (at least) two distinct kinds, or notions, of a person—let us call these $person_1$ and $person_2$—such that, in such cases, each of a and b is a $person_1$ and also a $person_2$, but because the two kinds differ in the identity conditions they specify for their members, a is (determinately) the same $person_1$ as b yet not the same $person_2$ (so that neither is essentially a $person_2$). This position, however, implies that $a = b$ and $a \neq b$. (The same inconsistency occurs in Johnston's 'Human Beings,' *Journal of Philosophy*, 84, 2 (February 1987), pp. 59–83, at 76. See also his 'Is There a Problem about Persistence?' at p. 123, bottom. Although Johnston opposes the Cartesian-dualist position that persons exist 'separately' from their bodies, his view that kinds specify identity conditions for their members, with different kinds specifying differing conditions, leads him to a position even more radical than the dualism he rejects: that we are not organisms, or indeed biological life forms of any kind.) A consistent variant would be this: a is both a $person_1$ and a $person_2$ at a time t_1, b is the same $person_1$, at a later time t_2 that a is at t_1 but b is *not* a $person_2$ at t_2 and consequently not the same $person_2$ at t_2 that a is at t_1. Since Johnston concedes that personal identity is strict identity restricted to persons, this alternative position reduces to the following: a is both a $person_1$ and a $person_2$ at t_1, whereas a is only a person, and no longer a $person_2$ at t_2. Whatever this prospect may mean for our ordinary concept of a person, it does not warrant the dramatic conclusion that the notion of personal identity is indeterminate for a. The alleged ambiguity may render some confusion over the issue of whether a is still a 'person,' but there is no lingering issue, and there should be no problem, concerning whether the thing at t_2 (whether or not it is still a person) is still the *same thing*, and if so, what makes it so. In the usual puzzle cases of the traditional problem of personal identity (including Johnston's favored puzzle case of fission), there is typically no serious question about the status of any of the relevant individuals as persons. Instead it is *given* that the principal individuals in question are persons. Typically, a is stipulated to be a person [man, woman] by hypothesis, while b is given descriptively as 'the *person* who emerges from such-and-such a process' (e.g., as 'the *a*-body-*person*,' or as 'the *man* who now has the left hemisphere of what was previously *a*'s brain,' etc.). The primary question concerns a's identity with, or distinctness from, b—not whether b is a person at t_2, or whether the erstwhile person a (whose identity with b is in question) is still a person at t_2. Indeed, the prospect that a is determinately no longer a person at t_2 (and for that reason alone, not the same person as b) is typically ignored altogether. (See note 29 below. Curiously, even Johnston does not consider this prospect in his cataloguing of potential solutions to the problem.)

conceptions of what a possible world is. (Not all of these need be thought of as competing conceptions.) The conception I favor is that of a maximally specific scenario that might have obtained.[13] On this conception (and on suitably closely related conceptions), each of the puzzle-case scenarios is the intersection of an infinite plurality of possible worlds, i.e. a constituent 'mini-world,' or sub-scenario, common to each. Each of the three puzzle-cases may be regarded as representing a distinct class of worlds. D/R-1, for example, represents the class of worlds in which Allen performs brain transplants with the result that the Rita-body-person is now an ideal mate and Allen has fallen in love with the Doris-body-person. The primary question for each of these scenarios is which identifications obtain in the worlds represented by that scenario.

Viewing the puzzle cases as representing classes of worlds, there appears to be some kinship between the problem of personal identity and another identity problem of contemporary philosophy: *the problem of trans-world identification*, i.e. the problem of identifying individuals in different possible worlds. Consider the possibility of Richard Nixon having continued as United States president for the duration of his second term in office. We may ask: Would the Democrats have regained the presidency, as they did in the actual world? Would they have nominated Jimmy Carter? And so on. But before we can answer, a philosopher interrupts. What determines whether the President in the possible world under discussion is Nixon? How can we know that it is Nixon rather than someone else who resembles Nixon in a variety of important respects, except for having finished out his presidency rather than resigning in disgrace? And furthermore, what does being Nixon consist in for someone in another possible world? In short, what is the *criterion*, or *criteria*, of trans-world identity that settles the question of whether someone in another possible world is Nixon? In a celebrated critique, Kripke has exposed the alleged problem of trans-world identity as a pseudo-problem (Kripke, pp. 15–20, 42–53, 76–77). He counters that possible worlds are not like independently existing planets with features to be investigated. ' "Possible worlds" are *stipulated*, not *discovered* by powerful telescopes,' he says. 'There is no reason why we cannot *stipulate* that, in talking about what would have happened to Nixon in a certain counterfactual situation, we are talking about what would have happened to *him*' (p. 44).

Kripke's contention that possible worlds are 'stipulated' has been seriously misunderstood.[14] Many philosophers take it as thesis a about the ontological and/or

[13] *Cf.* my 'The Logic of What Might Have Been,' cited above in note 4.

[14] A dramatic case in point is Allen Hazen, in 'Counterpart-theoretic Semantics for Modal Logic,' *Journal of Philosophy*, 76, 6 (June 1979), pp. 319–338. Hazen asserts (pp. 334–335) that when Kripke says that possible worlds are stipulated rather than discovered, what he means, in part, may be explained by saying that a possible world is a combination of a purely qualitatively specified world together with a particular stipulated choice among various similarity correspondences or mappings (which need not be one-one) between individuals in other worlds and individuals of the qualitatively specified world. Hazen thinks of the similarity correspondences as schemes that represent an individual in some other world by means of a selected counterpart in the qualitatively given world. Hazen's entire apparatus is decidedly anti-Kripkean. Kripke adamantly insists that possible worlds need not be purely qualitatively specified, and that the very same individuals may exist in different possible worlds rather than being represented in another world by 'counterparts' in that world.

epistemological status of possible worlds, about how they came into being and how we come to know of them. They see Kripke as a *modal conceptualist*, who believes that possible worlds are somehow created by us with the properties that we assign to them (a position analogous in certain respects to constructivism about mathematical entities). Readers have thought that Kripke holds that we are the masters of metaphysical modality, in the sense that it is entirely for us to decide, by 'stipulation,' what is metaphysically possible and what is not. These are serious misinterpretations. Kripke's observation that 'possible worlds are not discovered but stipulated' is simply his endorsement of a version of the doctrine that David Kaplan calls *Haecceitism*. The *haecceity* of an individual *x* is the property of being identical with *x*, i.e. the property of being *that very individual*. Kaplan defines Haecceitism as the doctrine that

> we can meaningfully ask whether a possible individual that exists in one possible world also exists in another without taking into account the attributes and behavior of the individuals that exist in the one world and making a comparison with the attributes and behavior of the individuals that exist in the other world . . . [the] doctrine that holds that it does make sense to ask—without reference to common attributes and behavior—whether *this* is the same individual in another possible world, that individuals can be extended in logical space (i.e., through possible worlds) in much the way we commonly regard them as being extended in physical space and time, and that a common 'thisness' may underlie extreme dissimilarity or distinct thisnesses may underlie great resemblance, . . . [15]

Despite the usual gloss on Kaplan's explanations, the central doctrine of Haecceitism is not concerned primarily with the identification of individuals in distinct possible worlds—although the doctrine does have important consequences concerning cross-world identifications. The central doctrine primarily concerns an issue of *legitimacy*. It concerns the question of whether it is 'meaningful' to stipulate the facts about particular individuals in particular possible worlds, including such facts as that the individual with such-and-such properties in a given world *w* is a particular individual *a*, or is not the particular individual *a*, as the case may be. Haecceitism holds that it is perfectly legitimate when introducing a possible world for consideration and discussion, to specify the world explicitly in terms of facts directly concerning particular individuals, designating those individuals directly by name if one chooses to.

An extreme version of the doctrine—*Extreme Haecceitism*, as I shall call it—combines Haecceitism in the preceding sense with a further doctrine: that facts concerning the particular individual *a* are in some relevant sense primitive, not reducible to any more general facts, such as that the individual with such-and-such properties is thus-and-so. Extreme Haecceitism holds that it is legitimate to stipulate facts concerning particular individuals in a world, identifying those individuals by name, precisely *because* such facts about a world are held to be *separate* facts that are not fixed by, and cannot be logically inferred from, facts that do not specify which individuals are involved. I shall use the term 'Reductionism' for the opposing doctrine that any such facts about a world *w* as that the individual with such-and-such

[15] Kaplan, 'How to Russell a Frege–Church,' *Journal of Philosophy*, 72 (1975), pp. 716–729, at 722–723.

properties is *a*, or is not *a*, if indeed such facts exist, are reducible to such qualitative facts as that the individual with such-and-such properties in world *w* is the individual with so-and-so properties in world *w′* (where the so-and-so properties are similar, or closely related, to the such-and-such properties).

Unfortunately, it is unclear what it means to say that facts of one kind are *reducible* to facts of another—or using alternative terminologies, that facts of the first kind 'consist in,' or are 'nothing over and above,' facts of the second kind, or that facts of the one kind are 'grounded in,' 'derived from,' 'based upon,' 'constructed out of,' or 'constituted by' facts of the other kind. The central idea seems to be that any fact of the first kind is a *logical* or *conceptual* consequence of facts of the second kind. An example would help enormously here. But there are precious few, if any, uncontroversial examples. One example from the philosophy of language may do. On Frege's philosophy of semantics, the referential (denotative, designative) facts concerning a language are reducible to other sorts of facts—in particular to intensional-semantic facts about what the sense of an expression is together with extra-linguistic facts about what a given sense *metaphysically determines*. To illustrate, the English noun 'water', in its use as a name for the familiar liquid, semantically expresses a certain concept (or property) *c* as its English sense, perhaps *the colorless, odorless, potable liquid found (with varying amounts of impurities) in lakes, rivers, and streams*.[16] This is a fact in the theory of meaning—a fact concerning the semantics of sense—and not a fact in the theory of reference. The concept *c*, in turn, metaphysically determines the chemical compound H_2O, in the sense that the compound exactly fits *c* and (let us suppose) no other substance does. This fact is completely independent of language. It is a straightforward logical consequence of these two—the meaning fact and the metaphysical fact—that there is some concept or other such that the word 'water' expresses that concept as its English sense and that concept in turn determines H_2O. The latter, according to a Fregean philosophy of semantics, just *is* the fact that 'water' refers in English to H_2O. This fact is thus partly semantic and partly metaphysical in nature.[17] In this sense, the fact that the English noun 'water' refers to H_2O is 'nothing over and above' (consists in, is grounded in, is derived from, etc.) the two facts that the English noun 'water' expresses *c* and that *c* determines H_2O.[18]

[16] I use the word 'concept' here in the same sense as Alonzo Church, which is decidedly distinct from that of Frege's artificial use of the German '*Begriff*'.

[17] In the terminology and conceptual apparatus of my 'Analyticity and Apriority,' in J. Tomberlin, ed., *Philosophical Perspectives, 7: Logic and Language* (Atascadero, Ca.: Ridgeview, 1993), pp. 125–133, the fact in question is (according to Frege's theory of it) a fact of *applied* rather than *pure* semantics, since it involves some extra-linguistic metaphysics.

[18] The notion of reducibility involved here will be clarified further in Section VI below. An alternative notion of reducibility results by replacing the relation of logical consequence with the notion (metaphor?) of part-whole constitution. We may say that a fact *f* is *mereologically reducible* to a class of facts *c* if *f* is literally composed, without remainder, of the elements of *c*. Thus a mereologically complex fact is mereologically reducible to its constituent sub-facts. This notion is suggested by a more literal construal of the terminology of one fact being nothing over and above, or consisting in, etc., a plurality of other facts. The notion presupposes a picture of compound facts as complex wholes resulting from an assemblage of other facts. This picture raises baffling questions about the relationship between mereological reducibility and the logical or conceptual notion of

A doctrine more extreme than simple Reductionism opposes simple Haecceitism. *Anti-Haecceitism* is the doctrine that in introducing a possible world for consideration and discussion, one may not legitimately specify facts while mentioning the individuals involved by name (or by something similar, such as by a demonstrative uttered while pointing to an actual individual). Instead, one may specify only the general, qualitative sorts of facts to which the facts concerning a particular individual (if there are any such facts) are reducible according to Reductionism. Specifying the facts concerning a particular individual *a*, explicitly identifying *a* by name, is regarded as a form of cheating—or rather, it is held to be meaningless. Some Anti-Haecceitists go so far as to reject the very existence of such facts about a world as that the individual with such-and-such properties is, or is not, the very individual *a*. They hold that one may not legitimately specify such facts in giving a possible world for the simple reason that there are no such facts to be specified. This view might be called 'Extreme Anti-Haecceitism'. Less extreme Anti-Haecceitists embrace Reductionism, holding that while there are facts directly concerning specific individuals, they are reducible to general facts to the effect that the individual with such-and-such properties is, or is not, the individual with so-and-so properties. Extreme Haecceitism, in contrast to Anti-Haecceitism, and in sharp contrast to Extreme Anti-Haecceitism, holds that the former facts are *further* facts over and above general facts, not reducible to or constructed out of the latter. Along with the general facts, these separate facts concerning specific individuals are held to be built into the very fabric of the possible worlds themselves.

Little or no notice has been made in the extant literature on Haecceitism of the distinction between the moderate and extreme versions of these various doctrines. I have endeavored to make my usage correspond as closely as possible to established usage of the terms 'Haecceitism' and 'Anti-Haecceitism'. That is why I have introduced the special terms, 'Extreme Haecceitism' and 'Reductionism', for the opposing doctrines concerning the question of reducibility (which is less often the primary focus), and a third term, 'Extreme Anti-Haecceitism', for what may be the most controversial of the doctrines. Extreme Haecceitism and Reductionism are

reducibility explicated in the text. On Frege's meta-semantical theory, is the fact that the English word 'water' refers to H_2O mereologically reducible to other facts? In particular, does it mereologically reduce to the pair of facts that 'water' expresses *c* and that *c* metaphysically determines H_2O? Is it supposed to be obvious that it does? Suppose 'water' had expressed a different concept in English, but one which also determines H_2O. Would the fact that 'water' refers in English to H_2O then be a different fact, consisting of different sub-facts? Let us say that the proposition that such-and-such, if it is true, *corresponds to* the fact that such-and-such. On some theories, this relation of correspondence is simply identity restricted to true propositions. Suppose that a proposition *p* corresponds to a mereologically reducible fact *f*, and that propositions $q_1, q_2, q_3, \ldots$ correspond to the sub-facts to which *f* mereologically reduces. Is *p* then logically equivalent to the conjunction (q_1 and q_2 and q_3 and . . .)? Or is *p* merely a logical consequence of the conjunction? Or might the two even be logically independent?

Lacking answers to these and other questions, I shall rely in the text primarily on the conceptual notion of reducibility that invokes logical consequence rather than the part-whole relation. It may be useful, however, to bear in mind the possibility that a particular author may instead mean the mereological notion, or something else. Where appropriate, one should distinguish between Mereological Reductionism and Conceptual Reductionism (the notion explicated in the text), as I shall do in some notes below.

the exact denials of one another. Extreme Haecceitism, therefore, might also be called 'Anti-Reductionism'. One may consistently combine Haecceitism (*simpliciter*) with Reductionism by holding that it is legitimate to introduce a possible world for consideration by stipulating which facts concerning particular individuals obtain in the world even though such facts are reducible to, or nothing over and above, other sorts of facts. (It is possible that Kripke takes this position. See note 24 below.)

The various versions of Haecceitism and Anti-Haecceitism are perhaps best formulated by invoking a concept from the theory of propositions, that of a *singular proposition*. A singular proposition is a proposition in which at least one individual or object that the proposition is about occurs directly as a constituent, and the proposition is about that individual by virtue of directly including it, rather than a concept by which the individual is represented (determined, denoted). In introducing the terminology of 'singular propositions', Kaplan equates Haecceitism with the acceptance of singular propositions (*ibid.*, pp. 724–725). More accurately, Haecceitism is the doctrine that one may legitimately cite singular propositions in specifying the propositions that are true in a possible world introduced for discussion. Extreme Haecceitism is the stronger doctrine that the truth values of any and all manner of singular propositions are among the primitive, brute facts about which propositions are true and which are false in a given possible world. If one conceives of possible worlds as maximal compossible sets of propositions, then Haecceitism holds that possible worlds include singular propositions among their elements in addition to non-singular, or general, propositions, and Extreme Haecceitism holds that the entire subset of non-singular propositions included in a world to the effect that the F is such-and-such, for particular properties F, logically entails no singular proposition to the effect that x is such-and-such. Reductionism holds that the subset of singular propositions, assuming one countenances such propositions at all, is fixed by the subset of non-singular propositions. Anti-Haecceitism (*simpliciter*) holds that possible worlds include *only* general propositions to begin with, leaving open the question of the truth values of any singular propositions, and Extreme Anti-Haecceitism denies that there are any singular propositions to be concerned about.

Kaplan points out that one should strictly speak of Haecceitism, Anti-Haecceitism, and their variants as relativized to a particular kind of entity K, as for example, *Anti-Haecceitism with regard to concrete things, Reductionism with regard to social institutions*, etc. Reductionism with regard to political nations, for example, is the often-cited doctrine that facts involving political nations are reducible to other sorts of facts, such as the actions and histories of particular persons. Extreme Haecceitism regarding political nations is the denial of this alleged reducibility. Haecceitism with regard to a kind K is logically independent of Haecceitism with regard to any logically independent kind K'. One may consistently combine Haecceitism regarding human bodies with Anti-Haecceitism regarding persons, for example, by holding that it is legitimate to specify which bodies exist in introducing a possible world for consideration but not to specify which persons exist in that world.

The astute reader will have noticed that I have described the various versions and variants of Haecceitism and Anti-Haecceitism without mentioning the alleged problem of trans-world identification, focusing instead on the role of facts

concerning specific individuals in presenting a possible world. How does the trans-world identity problem come in? On Anti-Haecceitism regarding individuals, possible worlds do not include specific individuals themselves. Instead they provide a structure and framework, given purely qualitatively, in which individuals are represented by means of individual concepts. It is not labelled which individual a given individual concept represents. For the Anti-Haecceitist, then, there is a special problem about how the individuals thus represented in distinct possible worlds are to be identified with, or distinguished from, one another. If identification is your game, some assembly is required. And all one has to go on are the individual concepts that represent the individuals. One thus needs criteria of trans-world identity. There is no like problem for the Haecceitist, since facts concerning specific individuals may be given directly in specifying the possible worlds under discussion. This is what Kripke means when he says that a possible world need not be given purely qualitatively. Haecceitism holds that facts concerning the haecceities—or in more ordinary parlance, the identities—of specific individuals may be taken as given in introducing a possible world for consideration, and Extreme Haecceitism holds that all facts concerning specific individuals are directly settled by the internal make-up of the possible worlds themselves. Possible worlds come already equipped with identification labels for the individuals that exist in them. No assembly is required, no identity criteria needed.

Kripke's assertion that possible worlds are not discovered but stipulated is a somewhat less felicitous way of stating what I take to be the central doctrine of Haecceitism *simpliciter*, or a closely related doctrine. Criteria for trans-world identity are to be replaced by stipulations. In fact, in this respect possible worlds are no different from anything else that might come under discussion. Suppose I say, 'Some cities have monuments made of marble,' as a prelude to saying something about some or all such cities. It would be silly (at best) for someone to object that while there are indeed marble monuments in *this* city (the city we are in), I must justify my claim that the monuments in the other cities I have in mind are really made of marble—instead of, say, some other material that was fashioned to look the way marble looks around here. I am discussing cities with marble monuments. I do not have to specify the relevant class of cities purely qualitatively and then provide a criterion for inter-city identity of material. I simply select the class of cities that I wish to discuss by specifying that they have monuments made of . . . , well, *marble*. Kripke contrasts possible worlds, which he says are stipulated, with planets, which are discovered. This may have given the wrong impression. Even independently existing planets may be stipulated in the sense that Kripke intends. One astronomer says to another, 'There are undoubtedly thousands of planets that, like Earth, have significant amounts of oxygen in their atmospheres. What is the temperature range for such a planet?' Suppose a philosopher who has been eavesdropping interrupts, 'Not so fast. How do you know, and what makes it true, that the atmospheric gas on the planet in question is oxygen, rather than some other element that superficially resembles oxygen? After all, you're not on that planet; you're in no position to send up a weather balloon or to conduct other atmospheric experiments. Are you supposing that, say, atomic number provides a criterion for interplanetary identity of elements? If so, why atomic number? Why not some other feature, like that of

having its source in the particular portion of ancient post-Big-Bang material from which our Earth-bound oxygen was originally formed?' A reaction by the astronomers of eye-rolling annoyance would be completely justified. The astronomer simply *stipulated* that he discussing planets that have significant amounts of oxygen in their atmospheres. Even if interplanetary identity criteria for elements are readily available, our astronomer is under no obligation to specify the planets he has in mind purely qualitatively and then ensure that they contain significant amounts of oxygen by providing the available criteria. It is in this sense that even planets are 'stipulated.' When Kripke says that we do not discover but stipulate possible worlds, he is not making a special claim about their peculiar ontological or epistemological status, or about our peculiar status *vis a vis* possible worlds. Nor is he claiming that we decree what is possible and what is not. Instead what he means is that the question of which class of possible worlds is under discussion (and in particular the question of which individuals exist in those worlds) is like the matter of which class of entities of any sort is under discussion—whether they be animals, vegetables, minerals, sticks, stones, or even planets. It is a matter that is entirely open to, and may be entirely governed by, the stipulations of the discussants. The possibility of simply stipulating which individuals are involved renders trans-world identity criteria unnecessary.

V

Does the debate about Haecceitism have any bearing on the problem of personal identity? The problems of trans-world identification and of personal identity differ from each other in at least one relevant respect. The personal-identity puzzle cases begin with the stipulation that Doris and Rita are present in each. There is no question of identifying the Doris of D/R-1 with the Doris of D/R-2 or the Doris of D/R-3. For one thing, we are given that it is the same Doris in each scenario. For another, that does not help. We are not attempting to identify individuals across possible worlds. Instead we are attempting to identify individuals within a possible world (or within each of the possible worlds represented by the scenario under discussion). Kripke's observation about the stipulatory character of cross-world identifications appears to offer little help.

This appearance is deceptive. We are attempting to determine the identity (haecceity) of the Doris-body-person in D/R-1. This may be thought of as an attempt to identify an individual in an arbitrary possible world w of type D/R-1 with an individual of a possible world w', where the former is given qualitatively by means of the individual concept *the woman who now occupies such-and-such body*, and the latter is given directly, i.e. haecceitally, as either Doris or Rita. It happens that $w = w'$. This may be regarded as a special limiting case of the problem of trans-world identification in which the worlds in question are identical. Seen in this light, it emerges that the issue of Reductionism and the controversy between Haecceitism and Anti-Haecceitism are relevant to the problem of personal identity.

One point about the traditional problem of personal identity is perhaps obvious to anyone familiar with the topic. The problem presupposes a version of Reductionism

regarding persons. It is safe to say that nearly all writers on the topic of personal identity are Reductionists. Nearly everything in the literature on the topic simply assumes Reductionism regarding persons without mentioning it as such.

It is therefore ironic that Reductionism regarding persons entails the Dorita hypothesis. This is shown by a variation of the proof given in Section III above that for any pair of objects x and y, there is a determinate, objective fact of the matter as to whether $x = y$. Consider D/R-1. Let us name the Doris-body-person in D/R-1 'Doris-bod$_1$'. Suppose first, for the sake of argument, that Doris-bod$_1$ is Rita rather than Doris. Reductionists who make this identification claim that the fact that Doris-bod$_1$ = Rita is grounded in the fact that Doris-bod$_1$, and no one else, now has exactly such-and-such a psychology, which used to be Rita's psychology before the brain transplant. If this hypothesis is correct, then it yields one respect in which Doris-bod$_1$ differs from Rita. For the fact that Rita = Rita is a fact of logic, grounded in her existence perhaps but not in facts about her psychological history. Rita therefore lacks Doris-bod$_1$'s property that the fact that she is Rita is reducible to in her psychological history. Conversely, Doris-bod$_1$ lacks Rita's property that the fact that she is Rita is independent of psychological features of Doris-bod$_1$'s biography. Either way, it follows by Leibniz's Law that Rita $\neq$ Doris-bod$_1$, contradicting our hypothesis. But the alternative hypothesis that Doris-bod$_1$ = Doris is subject to refutation by an exactly analogous argument, employing reducibility to facts about Doris-bod$_1$'s bodily history in lieu of reducibility to facts about her psychological history. Either way, whether it is judged that Doris-bod$_1$ is Rita or Doris, the Reductionist is driven, or at least committed, to giving up that judgment. And this leads to the Dorita hypothesis. An exactly similar argument may be made in connection with D/R-2 and D/R-3.

This is an uncomfortable result for Reductionists. Insofar as the Dorita hypothesis is regarded as implausible with regard to any of the puzzle-case scenarios, so to that same extent is the Reductionist assumption that personal identity is grounded in such matters as psychological or bodily continuity. Assuming that one or the other of the rival hypotheses is correct, the thesis that the haecceity of Doris-bod$_1$ is metaphysically reducible to other facts—facts about her psychological or alternatively facts about her bodily history—is thereby disproved.

In fact, a version of Extreme Haecceitism (Anti-Reductionism) is susceptible of a variation of the same proof. Suppose, for a *reductio*, that there is an object x from a possible world w and an object y from a possible world w' such that the fact that $x = y$ is reducible to (or consists in, is nothing over and above, is derived from, etc.) general facts about x in w and y in w'. Their identity might be reducible, for example, to x's bearing the relation R in w to the same F to which y bears R' in w', for appropriate intra-world relations R and R' and an appropriate cross-world sortal F. It is evident, by contrast, that the fact that $x = x$ is not similarly reducible to general facts about x in w or in w'. For the fact that $x = x$ is a fact of logic. If it is grounded in any other fact at all, it is grounded only in x's existence (in w or in w'). But then x differs from y in at least one respect. For x lacks y's feature that its identity with x is grounded in general (cross-world) facts about x and it. Conversely, y lacks x's feature that its identity with x is a primitive fact, not grounded in any general facts

about *x* other than its existence. Either way, it follows by Leibniz's Law that *x* and *y* are different objects, contradicting the hypothesis that they are identical.[19]

Can we simply *stipulate* that the Doris-body-person in D/R-1 is, say, Rita? Haecceitism regarding persons implies an affirmative answer. And indeed on Extreme Haecceitism regarding persons, the matter of whether the Doris-body-person is Doris or Rita *should* be stipulated, since the identity (haecceity) of the Doris-body-person is a *further* fact, not reducible to such qualitative facts as that the Doris-body-person now has such-and-such a psychology (formerly characteristic of Rita). If we can simply stipulate that the Doris-body-person is Rita, then we should be equally free to stipulate instead that the Doris-body-person is Doris. Again, Haecceitism regarding persons implies that this is indeed so. Of course, the Doris-body-person cannot be both Doris and Rita. But we are not considering making both stipulations simultaneously. We are considering selecting one of them. And why not?

There is no particular reason why not. We *can* legitimately do this. As we have seen, the particular scenario D/R-1 represents a class of worlds. That class, it turns out, is diverse. The primary question concerning D/R-1 presupposes that in each of the worlds represented by that scenario, the identifications go the same way. This presupposition is erroneous. In some of the worlds represented by D/R-1, the Doris-body-person is Rita. In others of those worlds, the Doris-body-person is Doris. It is illegitimate to ask whether the Doris-body-person in D/R-1 is Doris or Rita. This is a matter to be settled by a stipulation concerning which worlds of the D/R-1 type are under discussion. We may say, 'Consider a world of type D/R-1 in which Allen performs brain transplants on Doris and Rita with the result that they have exchanged bodies. In any such world, Allen thereby made Doris into an ideal mate, but fell in love with Rita.' We may also say, 'Consider another world of type D/R-1, different from the last one, in which again Allen performs brain transplants on Doris and Rita, only in this case their individual consciousnesses remain with their bodies, so that they have exchanged their brains and their psychologies. In any world of this alternative sort, Allen thereby made Rita into an ideal mate, but fell in love with Doris.' Given Extreme Haecceitism, both sorts of worlds—both of these scenarios—are equally legitimate. They are equally legitimate *qua* scenarios. Neither is incoherent.

When a philosopher poses the D/R-1 scenario (or the D/R-2 or the D/R-3 scenario), and asks whether the Doris-body-person is Doris or Rita, and how this is supposed to be determined, the Extreme Haecceitist response—what I believe to be the correct response—goes something like this: You tell us who the Doris-body-person is. Until you do, you have not provided a scenario that is specified fully enough to settle the question. In response to your meta-question(s), it is not for us to *determine* which way the identifications go. It is up to you to *stipulate* which class of scenarios you have in mind. As stated, your questions presuppose that the identifications automatically go the same way for all scenarios of the relevant type. Since the identifications you seek are not reducible to the facts you have given us, that presupposition is false. Until you make the necessary stipulations, your primary

[19] See my 'The Fact that $x = y$,' *Philosophia* (Israel), 17, 4 (December 1987), pp. 517–518. For a variety of controversial, but similarly proved philosophical theses concerning identity, see the appendix to my 'Modal Paradox,' pp. 110–114. (*Cf.* especially T6 and T7 listed there.)

question is unanswerable in principle. And once you make the necessary stipulations, the answer is then trivial.[20]

VI

Given Haecceitism regarding persons, or at least given its Extreme cousin, the traditional problem of personal identity does not get off the ground. Yet an alternative version of the problem obstinately remains. Imagine that Allen actually *does* perform the operation on Doris and Rita. Imagine this really happening. Imagine that Allen really does—right here and now—implant what had been Doris's brain in what had been Rita's body and conversely. The Rita-body-person is now an ideal mate. Allen has fallen in love with the Doris-body-person. Who now has Allen fallen in love with?

This is not in any way a matter to be settled by stipulation. Surely there already is some fact of the matter concerning the Doris-body-person's identity. And it is not subject to our control what that fact is. If she is Rita, that is not at all a result of my (or of our) stipulating that this should be so. No one has made any such stipulation, nor would it have the slightest effect on things if one did. Instead the Doris-body-person's identity with Rita—the fact that Doris and Rita have exchanged bodies—seems to be somehow a result of the way the surgery was performed, somehow a result of the fact that the Doris-body-person now has what used to be Rita's brain and consequently also what used to be Rita's psychology. The whole business of identity criteria being replaced by Kripkean stipulations seems beside the point, if not completely wide of the mark.

One may feel uneasy about the idea of going beyond mere consideration of the possibility of a given situation, and instead imagining it to be actual. We know it is not actual. Why pretend that it is?

For a simple reason. The point is to mobilize intuitions concerning what *would* be the case if D/R-1 *had* occurred. If, counterfactually, Allen had performed brain transplants on Doris and Rita, then there would be a resulting fact as to whether the Doris-body-person was Rita or Doris, and that fact would not be a matter of our stipulating what is so. Kripke's observation that 'possible worlds are stipulated,' properly understood, is simply a recognition of the fact that in considering certain possibilities, we are free to stipulate which possibilities we have in mind by specifying which individuals are involved in them. As we have already seen, it is not a thesis to the effect that what is possible with respect to those individuals is subject to our decision. Nor is it a thesis to the effect that we decide what *would* be the case under certain counterfactual circumstances. There is already a fact of the matter, independently of us, as to who the Doris-body-person would be if D/R-1 had occurred.

Let us suppose again that the Doris-body-person would be Rita. If this hypothesis is correct, it appears to be a direct result of the fact that the Doris-body-person has what was previously Rita's brain with Rita's psychology relatively intact. Insofar as it is true that if D/R-1 had occurred, the Doris-body-person would be Rita, something

[20] *Cf. Reference and Essence*, at pp. 242–243.

significantly stronger is equally true. It is not as if the D/R-1 scenario might have had different results. If the Doris-body-person would have been Rita had D/R-1 occurred, then it is in fact metaphysically *impossible* for D/R-1 to occur with the Doris-body-person being Doris, or anyone else other than Rita. In a word, it is *necessary* that the Doris-body-person in D/R-1 is Rita.

Earlier I said that the class of worlds represented by the D/R-1 scenario was diverse, that there are possible worlds in which the D/R-1 scenario is realized and the Doris-body-person is Rita and other worlds in which the D/R-1 scenario is realized and the Doris-body-person is Doris. Now I am saying that the latter outcome is impossible, that there are no possible worlds in which the Doris-body-person is Doris. I seem to have contradicted myself.

I have not. It is at this juncture that I invoke impossible worlds. Haecceitism does not entail that it is in some way for us to decide what is, and what is not, metaphysically possible. Even Extreme Haecceitism. does not entail this. Haecceitism simply holds that in introducing a world for consideration and discussion, we are free to stipulate the facts that obtain in the world. Depending on what we stipulate, the world, or worlds, we so introduce may turn out to be impossible rather than possible. This is so even if it was our intent to stipulate a possible world. We decide which individuals exist and what properties they have in the world we wish to consider, but Metaphysics decides, under its own authority, whether such a world is possible or impossible. The latter issue is completely out of our hands. There are indeed D/R-1 worlds in which the Doris-body-person is Rita, and there are indeed other D/R-1 worlds in which the Doris-body-person is Doris. For that matter, there are D/R-1 worlds in which the Doris-body-person is Madonna (altered to have Rita's psychology), and still other D/R-1 worlds in which the Doris-body-person is Ethel Merman resurrected (and psychologically altered). This is a consequence of Extreme Haecceitism. The question of the Doris-body-person's haecceity—the question of who the Doris-body-person is—is not to be found among, and does not reduce to or consist in, the facts that are given in the D/R-1 scenario. There are many different ways for the identifications to go. But most of those ways are quite impossible. In all of the genuinely *possible* D/R-1 worlds, the Doris-body-person is Rita. This is fixed by law but not by legislation. It is fixed by Metaphysical law.

It emerges from this analysis that there are two very different ways of interpreting the problem of personal identity, depending on whether Reductionism is presupposed. A puzzle case like D/R-1 is first set out, and the primary question and the two meta-questions then posed. If the questions are put forward under the presupposition of Reductionism, it is assumed that one has been given all the facts that are required for deciding the primary question, taken as a question about *all* the worlds represented by the puzzle-case scenario, possible and impossible. One may restrict one's focus to possible worlds, but there is no need to do so. The same answer will obtain for the impossible worlds as well, or at least for the logically consistent ones. For the Reductionist, so-called criteria of identity are reductionist analyses or definitions of what it is for a pair of individuals at different times or in different worlds to be identical—or at least analytic sufficient conditions for

cross-circumstantial identity. The metaphysical meta-question is concerned with the presupposed reduction of personal-identity facts to facts about psychologies or about bodies. It is, in effect, a demand to be given a reductionist analysis for personal identity through change. We may call this *the Reductionist problem of personal identity*. It is the orthodox or canonical form of the problem.[21] As an Extreme Haecceitist, I reject this alleged problem as bogus (along with the alleged problem of trans-world identification).

If the primary question and the two meta-questions are put forward without presupposing Reductionism, one is then presumably being asked to confine one's attention to genuinely possible worlds. In those *possible* worlds in which D/R-1 (or D/R-2 or D/R-3) is realized, who is the Doris-body-person? In particular, if D/R-1 *were* realized, who *would* the Doris-body-person be? This question is perfectly legitimate. The facts of that case are sufficient to zero in on one metaphysically necessary outcome. That is to say, even if the Doris-body-person's identity (haecceity) is not *reducible* to the sorts of facts that one is given in D/R-1, the Doris-body-person's identity does *supervene modally* on exactly such facts. For present purposes, the relevant notion of supervenience may be defined as follows:

> Properties of kind K *modally supervene on* properties of kind $K' =_{\text{def}}$ For any class c of K-properties and for any class c' of K'-properties, if it is metaphysically possible for there to be something whose K-properties are exactly those in c and whose K'-properties are exactly those in c', then it is metaphysically necessary that anything whose K'-properties are exactly those in c' is such that its (his/her) K-properties are exactly those in c.

Thus, to say that K-properties modally supervene on K'-properties is to say that either it is metaphysically necessary that anything that has exactly such-and-such K'-properties also has exactly so-and-so K-properties or else it is metaphysically impossible for anything to have exactly such-and-such K'-properties and also have exactly so-and-so K-properties. Or put another way, which K-properties a thing has is metaphysically necessitated by which K'-properties it has. For example, to say that a person's psychology modally supervenes on his/her brain and its physical states is to say that a complete accounting of the facts concerning a person's brain and its physical states leaves room for only one possible outcome concerning his/her psychology, in the sense that it would be metaphysically impossible for the person's brain to be in exactly those physical states while the person has a different psychology (even one that is only slightly different). What I am claiming here is that

[21] I have borrowed the terms 'Reductionism' and 'further facts' from Parfit, who explicitly calls himself a 'Reductionist' in rejecting the idea that identity facts are further facts (*Reasons and Persons*, at p. 255). In 'Are Persons Bodies?', in his *Problems of the Self*, pp. 64–81, Williams defends his setting out the problem of personal identity by means of the BW device thus: 'Such a process may, perhaps, be forever impossible, but it does not seem to present any purely logical or conceptual difficulty' (p. 79). The exact intent of these remarks is perhaps unclear, but on one natural interpretation, Williams is prepared to allow for the prospect (putting the matter in terms of my apparatus) that all of the logically or conceptually possible worlds in which the D/R-2 scenario occurs are metaphysically impossible. Never mind; there is still supposed to be a problem. On this interpretation, the resulting 'problem of personal identity' is a problem only on the assumption of Reductionism.

the Doris-body-person's haecceity modally supervenes on, but is not reducible to, exactly the sorts of biographical facts given in D/R-1.[22]

One may define a notion of reducibility by means of a simple adjustment in the above definition of supervenience, changing the metaphysical modalities to *conceptual* (or properly *logical*) modalities. It may be assumed here that conceptual necessity entails metaphysical necessity but not vice versa. What is conceptually necessary is true in every conceptually possible world, including such worlds as are metaphysically impossible. To say, then, that properties of kind *K* are *conceptually reducible to* properties of kind *K′* is to say that for any class *c* of *K*-properties and for any class *c′* of *K′*-properties, if it is conceptually (or logically) possible for there to be something whose *K*-properties are exactly those in *c* and whose *K′*-properties are exactly those in *c′*, then it is conceptually (logically) necessary that anything whose *K′*-properties are exactly those in *c′* is such that its (his/her) *K*-properties are exactly those in *c*. The idea here is that either it is conceptually necessary (a logical or analytic truth) that anything that has exactly such-and-such *K′*-properties also has exactly so-and-so *K*-properties or else it is conceptually incoherent (logically inconsistent) for anything to have exactly such-and-such *K′*-properties and also have exactly so-and-so *K*-properties. Or put another way, which *K*-properties a thing has is a logical consequence of which *K′*-properties it has. For example, on Frege's meta-semantical theory, the referential semantics for a language is reducible to the language's intensional semantics (i.e., its semantics of sense) together with some metaphysics, in that the referential properties of a language are reducible to the language's sense properties taken together with the extra-linguistic matter of what objects are determined by those senses. Given that conceptual necessity entails metaphysical necessity but not vice versa, it follows that conceptual reducibility entails modal supervenience but not vice versa.[23] A claim to the effect that *K*-properties supervene on *K′*-properties therefore normally carries the implicature that *K*-properties are *not* reducible to *K′*-properties. And indeed, when philosophers explicitly advocate a supervenience thesis, they often explicitly contrast that thesis with the corresponding reducibility thesis, which they reject, or at least decline to endorse. I am doing exactly that here.

On the modal-supervenience interpretation of the problem of personal identity, the two meta-questions about 'criteria for personal identity' are distinct. The

[22] Jaegwon Kim defines some non-equivalent notions of supervenience in 'Concepts of Supervenience,' *Philosophy and Phenomenological Research*, 65 (1984), pp. 257–270. The notion defined in the text corresponds to Kim's favored notion of *strong supervenience* (where the modality involved is metaphysical modality).

If I am correct, recognition of the distinction between supervenience and reducibility is crucial if we are to make significant progress toward solving the traditional problem of personal identity. The Reductionist regarding personal identity typically supposes that the haecceity of the Doris-body-person ought to be not merely supervenient on the sorts of facts about her that are given in D/R-3, but reducible to them. The weaker doctrine that personal identity modally supervenes on, but is not reducible to, such biographical features of a person as his/her psychological or bodily history may be what Parfit means when he speaks of what he calls *the Further Fact View* (p. 210). (Presently I shall deny that the haecceity of the Doris-body-person even modally supervenes on the facts given in D/R-3.)

[23] Given a certain kind of mereological essentialism, it follows that mereological reducibility of the sort described in note 18 above likewise entails modal supervenience but not vice versa.

metaphysical question is the deeper of the two—or at least, the more metaphysical. It is a demand for a metaphysical principle, or principles, that entail the answer to the primary question. It is, in effect, a demand for an individual person *a*'s *essence*, in the sense of a property such that it is metaphysically necessary that someone has the property if and only if he or she is the very individual *a* and no other. Or perhaps it is a demand merely for a modally *sufficient* property for *a*'s haecceity, i.e. a property such that necessarily, anyone with that property is the very individual *a* and no other. Or at the very least, it is a request for an *essential* property of *a*, i.e. a property that *a* has necessarily. The sought-after modal property must be adequate to the task of answering the primary question, interpreted now as a question about genuinely possible worlds in which the puzzle-case scenario obtains. This is *the Essentialist problem of personal identity*, to be distinguished from the Reductionist problem. The Essentialist problem does not presuppose that the sort of fact sought in answer to the primary question is reducible to, or is nothing over and above, facts of some other sort. The problem is perfectly compatible with the Extreme Haecceitist thesis that identity facts are further facts. Even by the Extreme Haecceitist's lights, it may be seen as a legitimate, and nontrivial, philosophical problem.[24]

We have seen that modal supervenience differs from reducibility (in one sense) over the type of modality involved. The two interpretations of the problem of personal identity carry with them correspondingly different notions of necessity that are involved in the explication given in Section I above of the concept of a criterion of personal identity. We said that a criterion of personal identity was a triple consisting of a sortal property F and a pair of binary relations R and R', other than personal identity itself, such that it is somehow necessary that x is the same person at t that y is at t' if (or perhaps iff) there is some trans-temporal link of sort F to which x bears R at t and to which y bears R' at t'. A purely epistemological criterion emerges by taking the necessity involved to be epistemic, e.g., knowability *a priori*, or perhaps the weaker notion: *given what we know, it must be that* (i.e., the dual of epistemic possibility: *for all we know, it may be that*). This would answer the epistemological meta-question. The Essentialist problem of personal identity takes the necessity involved to be metaphysical necessity, i.e. truth in all metaphysically possible worlds.

[24] Kripke (pp. 50–53) describes a version of the problem of trans-world identification that he finds legitimate, adding explicitly (p. 51) that there is a similarly legitimate problem concerning identity over time. The alleged problem is concerned with identifying physical objects in different possible worlds given only the facts concerning the relevant molecules (or other, more basic components). Insofar as Kripke is distinguishing between a pseudo-problem of cross-circumstantial identification that presupposes Reductionism with a genuine problem that instead presupposes mere modal supervenience, I am here echoing his sentiments specifically in regard to the traditional problem of personal identity. The textual evidence inconclusively suggests, however, that Kripke's remarks concern the Reductionist problem (which I dismiss as bogus), as opposed to the Extreme-Haecceitist/Essentialist problem. See my 'Trans-World Identification and Stipulation.'

It is possible that Kripke endorses a Mereological Reductionism of the sort described in note 18 above, and that his problem of trans-world identification presupposes this kind of Reductionism rather than Conceptual Reductionism. Although Kripke advocates Haecceitism in its moderate form, discussions I have had with him (subsequent to the appearance of *ami ng and ec essity*) make me doubtful whether he is prepared to hold, as I do, that haecceities are separate from, or facts over and above, such facts about individuals as their molecular composition (though he may be). *Cf. ibid.*, at p. 51n; and my 'The Logic of What Might Have Been,' at p. 20n.

The Reductionist problem of personal identity takes the necessity involved to be truth in all logically possible worlds, whether metaphysically possible or metaphysically impossible. (The phrase 'criterion of identity' may not be entirely appropriate on the Essentialist interpretation of the problem, since it seems to carry with it in connotation an acceptance of the Reductionist construal. But I shall continue to use it.)

The literature on personal identity has suffered from a failure to distinguish sharply between the Reductionist and the Essentialist interpretations of the problem. Philosopher *A* provisionally proposes a solution that is (or that at least might be) appropriate to the Essentialist problem, only to have it dismissed by philosopher *B*, noting that the proposed criterion does not work for every conceivable case, and thus construing it as a solution to the Reductionist problem. It even happens sometimes that *A* and *B* themselves bear the relation of personal identity.[25] When the distinction between the two interpretations is not emphasized, there is also the opposite danger that a Haecceitist who rejects the (*sic*.) problem of personal identity as unreasonably demanding, construing it Reductionistically, will miss the significance of the Essentialist problem.

VII

Let us reconsider the primary question and the metaphysical meta-question concerning D/R-1, interpreted now as concerning the class of possible worlds (excluding the impossible worlds) incorporating that scenario. At the end of Section II, it seemed as though the psychology-based identifications were correct—or at least that the body-based identifications were clearly incorrect. We may now go further. It is

[25] I take Johnston's 'Human Beings' to be an example of the converse situation. Johnston sees the problem of personal identity in the standard Reductionist way. (His Reductionism regarding persons is evidenced by his emphasis in 'Fission and the Facts' on conceptual possibility and conceptual necessity, and by his use of such phrases as 'that in which personal identity consists' in 'Human Beings' and 'the core relations that actually constitute personal identity' throughout 'Reasons and Reductionism.' Unlike the typical Reductionist, though, Johnston does not claim that the haecceity of the Doris-body-person is reducible to the sorts of facts given in D/R-3. See note 22 above.) Frustrated by an alleged conflict of intuitions regarding scenarios like D/R-2 and by the failure of previous attempts to solve the problem of personal identity, Johnston concludes, erroneously in my view, that the standard philosophical methodology of putting hypothetical cases to the test of intuition is somehow misguided. He argues that one should address the metaphysical meta-question instead with an eye to the epistemological meta-question. This procedure may make sense from the Reductionist standpoint, since whatever else identity facts are, they are knowable. Johnston opts for a solution to the metaphysical meta-question which, while it may be appropriate for the Essentialist problem, would be proved mistaken from the Reductionist standpoint by the questioned method of testing cases against intuition. Johnston's failure to distinguish between the Reductionist and Essentialist interpretations of the problem is further evidenced by his complaint that, according to the challenged methodology, 'the supposition that I could survive my body's petrification implies that the relations that tie me to my body are contingent' ('Human Beings,' p. 71). The phrase 'supposition that I could' here means conceptual possibility, while 'contingent' evidently means metaphysical contingency. Johnston also conflates reducibility and mere supervenience in 'Reasons and Reductionism,' at pp. 590–591. See also his 'Fission and the Facts,' at p. 381.

evident that, necessarily, in D/R-1 the psychology-based identifications are indeed correct. There are D/R-1 worlds in which the Doris-body-person is Doris, but such worlds are one and all impossible. In every *possible* world in which D/R-1 occurs, Doris and Rita have simply traded bodies (apart from their brains). To this extent, our original intuitions about this case are correct.

It does not follow that some psychology-based criterion for personal identity (such as the memory-based criterion) yields a correct answer to our metaphysical meta-question, interpreted on the Essentialist scheme. Psychology-based criteria are not the only criteria according to which Doris and Rita have exchanged bodies in D/R-1. A brain-based criterion would issue the same identifications. A brain-based criterion usually echoes the psychology-based criteria in the identifications it makes, but there is divergence in cases of brain damage. If a person's brain is damaged to an extent that significantly affects his/her psychology—such as by significantly altering his/her personality and/or memories of past events—corresponding psychology-based criteria deem the resulting person to be numerically distinct from the person prior to the brain damage. If the brain nevertheless continues to function sufficiently to produce consciousness and a psychology adequate for being a person, the brain-based criterion judges the resulting person to be literally and numerically the same as the person before the damage—only now not the same as he/she used to be.

Given what science informs us about the importance of the brain to consciousness, there does not seem to be much room for debate. The brain-based criterion, construed as an Essentialist criterion for personal identity, is intrinsically more plausible than either the body-based or the psychology-based criteria. I am not my body. But neither am I my psychology *as such*—my thoughts, my personality, my memories, my beliefs. I am more closely bound to my *consciousness* than to any of these other things. Not to my 'stream of consciousness,' mind you—the flow of thoughts, feelings, sensations, experiences, etc.—but the consciousness itself, the arena through which the flow flows. I may not be strictly identical with my consciousness. I continue to exist even through periods of unconsciousness (e.g. when asleep), even if not through all such periods. But there seems to be some connection between my consciousness and myself that is more intimate than that between my body and myself.[26] The brain is the organ that produces consciousness. Perhaps no one can say exactly how the brain does it. It may be that, at some sufficiently deep level of understanding, it is *impossible* to know how the brain does it. But somehow the brain does it, and that is something we do know. This knowledge provides forceful intuitive support for a brain-based Essentialist criterion for personal identity.

The body-based and psychology-based criteria each yield the same identifications in D/R-2 as they did in D/R-1. But the brain-based criterion has a special problem with D/R-2. Here the Doris-body-person has what was previously Doris's brain as well as what was previously Doris's body, but her brain now holds the information that was extracted from Rita's brain. Even if one has decided to make the identifications by attending to the brains rather than to the bodies or the psychologies, one

[26] The identification of a person with a consciousness, as opposed to a stream of consciousness, probably lies behind Descartes's proof of his own existence *via* his '*Cogito ergo sum*'. Ironically, it also lies behind Hume's denial of his own existence.

still has to decide whether the person's identity goes with the brain itself or instead with the information held within the brain. In D/R-2 these two come apart. The brain-*qua*-organ-based criterion is obtained by letting the sortal *F* in our explication of a personal-identity criterion be *brain* and letting both *R* and *R′* be the relation of *being the functional owner of*—which in this case may be taken to be the relation *u is the person whose brain is v*. The brain-information-based criterion is obtained instead by letting the sortal *F* be *brain-information* (operating system and RAM, etc.) and letting both *R* and *R′* be the relation: *u is a person whose brain holds exactly the information v*. According to the brain-*qua*-organ-based criterion, *x* is the same person as *y* if they have the same brain across time. According to the brain-information-based criterion, *x* is the same person as *y* if their brains store the same information across time. The brain-*qua*-organ-based criterion goes with the body-based criterion in D/R-2, the brain-information-based criterion with the psychology-based criteria.

Williams says that the primary advantage of setting out the problem of personal identity by means of the BW device rather than by means of brain transplants comes from the fact that D/R-2 is less radical than D/R-1 in the way it secures the condition that the Doris-body-person is appropriately connected to Rita so that the Doris-body-person's apparent memories of Rita's past are not automatically disqualified from being genuine memories. This remark of Williams' strongly suggests—and indeed much of the literature on personal identity assumes—that the identifications should come out exactly the same in both D/R-1 and D/R-2. (See note 11 above.) But it is at least potentially a mistake to assume this at the outset, without any argument or further ado. This is especially true since D/R-2 forcefully challenges the psychology-based criteria that seem so fitting when considering only D/R-1. It is logically possible, for example, that although the Doris-body-person is Rita in D/R-1, the Doris-body-person is instead Doris in D/R-2. This would have the consequence that the Doris-body-person of D/R-1 is numerically distinct from the Doris-body-person of D/R-2. But this is a logical possibility.[27]

I believe this logical possibility is philosophical reality. The primary question not only about D/R-1, but equally about D/R-2, when interpreted on the Essentialist scheme, is legitimate. In any genuinely possible D/R-1 world, the Doris-body-person is Rita, owing to the fact that the Doris-body-person in D/R-1 has what had been Rita's brain, still functioning in a normal manner. By contrast, in any possible D/R-2 world the Doris-body-person is Doris. And this is a result of the fact that the Doris-body-person in D/R-2 has what had been Doris's brain, still functioning in a normal manner. One noteworthy feature of the brain-*qua*-organ criterion—and an important argument in its favor—is that it discriminates between D/R-1 and D/R-2. It does indeed seem possible for a person to be given a different body by transplanting his/her brain into it. And it seems equally possible for a person to have his/her psychology radically altered by inducing substantial changes in his/her brain.

[27] The expressions 'the Doris-body-person' and 'the Rita-body-person' are definite descriptions (where a *Doris-body-person* is defined as being a person who now occupies what was Doris's body before the relevant procedure). There is therefore nothing about their semantics, as such, that requires them to be rigid designators, in the sense of Kripke.

Indeed, it is not an uncommon occurrence for someone's psychology to become significantly altered with brain damage. What seems impossible is for a person to take possession of a new body merely by having his/her psychology replicated in the new body's brain while his/her old brain is destroyed, or for a person's psychology to be modified by transplanting someone else's brain, with its readymade psychology, into his/her body. Of the several personal-identity criteria considered so far, the brain-*qua*-organ-based criterion is the only one that captures all of these intuitions. If one views the Essentialist problem of personal identity as a multi-partied election, then at this stage of the campaign at least, given our current state of knowledge, the brain-*qua*-organ-based criterion probably deserves one's vote. One thing seems clear: the rival criteria do not.

The solution I favor for the Essentialist problem of personal identity is to look neither to the body nor to the psychology, but to the organ of consciousness: the brain. I tentatively submit a pair of modal principles concerning persons and their brains. As a first approximation, consider the following essentialist principle:

> Necessarily, for any person P and any person-brain B, if B is P's brain at some time t (i.e. if P's consciousness is produced at t by B, etc.), then necessarily, for any time t' at which P is not brain-dead, P's brain at t' is B, so that if P is conscious at t', his/her consciousness is produced at t' by B.

The idea is that a person's brain is an essential property of the person, in the sense that as long as he or she is not brain-dead, his or her brain must be *that very brain* and no other. A different person-brain (for example, an artificial brain), no matter how extensively it replicates a person's original brain, cannot take the place of the original brain for that person. If the new brain produces consciousness, it is not that person's consciousness. Let us call this principle *the essentiality of one's brain*. The principle does not entail that a person's brain cannot undergo change. A person's brain might become damaged or undergo various surgical improvements (e.g., removal of tumors). The principle does not even deny that parts of a brain might be replaced with artificial components. What the principle entails is that whatever changes a brain undergoes, it must remain the same, numerically identical brain if its functional owner is to remain the same, numerically identical person. Replacement of a functional brain is homicide.

Kripke and others have proposed other essentialist principles concerning individuals, e.g. that the original material out of which an artifact was constructed is an essential feature of the artifact, and that any natural kind (e.g. the species) to which a creature belongs is an essential feature of the creature. As with these other principles, the essentiality of one's brain is *a posteriori*. It is subject to falsification and adjustment by the empirical facts. And as with other *a posteriori* essentialist principles, there is some more general, *a priori* essentialist principle from which the essentiality of one's brain is obtained. This may be the principle that if there is a single organ that is responsible for a person's consciousness, then it is essential to the person that he/she have that very organ (and not, for instance, a transplanted organ of the same type from someone else). The essentiality of one's brain derives its aposteriority from that of the supplementary observation that the organ of consciousness among persons is

the brain. The latter observation is subject to falsification by the improbable empirical discovery of a person lacking a brain.

The general principle itself is also subject to revision through critical inquiry, perhaps even through empirical findings. Philosophers often ponder the prospect of fission, whereby a single person is divided into two people by bisecting his/her brain and transplanting one or both of the brain-halves into a brainless body. It may be that anything that might be reasonably called a *person* must, as a matter of actual physiology, have more than one-half of an ordinary brain.[28] The loss of a smaller portion of one's brain might in some cases be regarded as damage that the brain survives—that is, as the brain's losing a part without thereby ceasing to exist as a functional brain. But if it should turn out that (as is frequently supposed in the relevant literature) enough of a person's psychology and consciousness can be retained with only one-half (or even somewhat less) of a brain, we may decide to replace the principle of the essentiality of one's brain with a weaker principle of *the essentiality of some sufficient portion or other of one's brain*. According to this essentialist principle, a person could survive the destruction of his/her brain by retaining a sufficiently functional portion. An Essentialist version of the fission problem might then arise.[29] It is difficult to conjecture about what the limitations are. Perhaps gradual bionicization is a real possibility. Perhaps different brain functions, including different aspects of consciousness, can be gradually taken over by different artificial devices, making the brain itself dispensable, thus requiring further modification in the essentiality of one's brain. But suppose the discarded brain were refurbished. Who would its functional owner be?[30]

If it is assumed that Doris and Rita continue to exist in both D/R-1 and D/R-2 after the relevant procedure, then the essentiality of one's brain answers the primary questions, as interpreted on the Essentialist scheme. Given Doris's and Rita's survival, the principle entails that they have exchanged bodies in D/R-1 and have

[28] So argues John Robinson in 'Personal Identity and Survival,' *Journal of Philosophy*, 85 (1988), pp. 319–328.

[29] The fission problem is analogous in some respects to a similar problem concerning artifacts, as illustrated by the famous Ship of Theseus. The former problem, interpreted on the Essentialist scheme, may be amenable to an analogue of the solution I proposed to the latter in *Reference and Essence*, pp. 219–229. There are alternative solutions to the fission problem which are not usually considered but which, if sound, would save the principle of the essentiality of one's brain. One is the claim that what survives the removal of a brain hemisphere is not the original person, but only what had been a part of that person and what is now a full-fledged (though perhaps impaired) person in his/her own right. In this case, the two persons who emerge from fission were formerly not persons at all, but two halves of the original person, who was destroyed. A variant of this solution holds that the original person continues to exist even after the fission, but only as the scattered aggregate of two separate persons, and therefore not itself a person. The fission would in that case constitute a radical metamorphosis whereby what had been a person is transformed from a solo act into a duo. On both of these solutions, any person who loses one of his/her brain's hemispheres is distinct from the person (or from each of the persons) who emerges from the procedure with only the remaining hemisphere. Both of these prospects are worthy of more serious attention.

[30] One carefully guarded variation of the principle in the text is the following: Necessarily, for any person P and any person-brain B, if B is P's brain at some time t, then necessarily, for any later time t' at which P is not brain-dead and such that P has a functioning brain throughout the period from t to t', P's brain at t' is a portion of B.

exchanged psychologies in D/R-2. In both cases they retain their individual brains. Indeed, according to the principle, they *must* retain their individual brains if they are to survive the relevant procedure, whether it is a brain transplant or a BW-exchange. This suggests an complementary essentialist principle:

> Necessarily, for any person-brain B and any person P, if B is P's brain at some time t, then necessarily, for any time t' at which B is functioning roughly normally, the person whose brain at t' is B is P, so that if B is producing consciousness at t', the consciousness produced is P's.

The idea is that a brain's functional owner is an essential property of the brain, in the sense that, as long as the brain is functioning in a substantially normally manner (allowing for some malfunctioning due to brain damage, etc.), the brain's functional owner must be *that very person* and no other. (It is assumed that necessarily, for any time t, any person-brain that is functioning substantially normally has exactly one functional owner at t, i.e. there is exactly one person whose brain it is at t.) No matter how much the psychology may have been altered—due to brain washing, under the influence of drugs or religious fanaticism, etc.—if the same brain is still producing consciousness in a more-or-less normal manner, it is the same person, even if he/she has been psychologically deeply altered in the process. We may call this principle *the essentiality of a brain's ownership*.[31]

Seen in one light, the twin principles of the essentiality of one's brain and the essentiality of a brain's ownership are the same but for a different focus. The two principles may be combined into a single principle of the *essentiality of brain ownership:*

> Necessarily, for any person P and any person-brain B, if B is P's brain at some time t, then necessarily, for any time t' at which either P is not brain-dead or B is functioning substantially normally, P is the person whose brain at t' is B.

A couple of points bear repeating here. First, the issue of whether this principle is correct is one that is appropriately settled partly by reference to empirical facts and partly by philosophical inquiry. Second, whether it is the essentiality of brain ownership or some alternative essentialist principle that is supported by empirical facts and philosophical analysis, the resulting 'criterion' for personal identity solves the Essentialist problem, not the Reductionist problem. The necessity involved in any brain-based criterion cannot be conceptual or logical necessity. It is manifestly not conceptually or logically necessary (e.g., it is clearly not an analytic truth) that persons have brains at all, let alone that a person has the same brain as long as he/she has the capacity for consciousness. Just as it is logically possible for a tin man to lack a heart yet live, it is likewise logically possible for a brainless scarecrow to be magically conscious, even impressively clever.[32]

[31] By the definition of supervenience given in Section VI, if one's haecceity modally supervenes on the original ownership of one's current brain, it follows that for any person x, no one other than x can possibly currently have what was originally x's brain. This is, in effect, the principle of the essentiality of a brain's ownership.

[32] *Cf.* David Wiggins, *Identity and Spatio-Temporal Continuity* (Oxford: Basil Blackwell, 1967), at p. 55. Williams evidently denies the principle of the essentiality of one's brain. In 'Are Persons Bodies?', he asserts that 'it seems pretty clear that under these circumstances [in which the BW device is used to copy information extracted from one brain into another] a person could be

An essentialist principle is a principle of modal intolerance; it imposes limitations on the variety of genuinely possible worlds. The essentiality of brain ownership does not concern impossible worlds. I am not proposing that a person's identity is *reducible* to (or that it is nothing over and above, or that it consists in, or is grounded in, derived from, etc.) facts about brains and their former owners. I maintain that the matter of the haecceity of a person given qualitatively as *the person who now functionally owns brain B* is a further fact. The traditional, canonical form of the problem of personal identity is correctly solved by rejecting it as a spurious pseudo-problem. One is free to stipulate that one is considering worlds in which the person who now has what used to be *a*'s brain is not *a* but someone else. There is a price to be paid for doing so: the worlds under consideration will be metaphysically impossible worlds. But there is absolutely no problem with that.[33]

counted the same if this were done to him, and in the process he were given a new brain . . . here we have personal identity without the same brain, though of course we have identity of the rest of the body to hold onto' (*Problems of the Self*, p. 80). Parfit argues similarly in *Reasons and Persons* (see note 10 above) that retaining some or all of one's brain is not what fundamentally 'matters' in survival. Although Williams may be a Reductionist (see note 21 above), his claim here is framed using the subjunctive construction 'if this *were* done to him, . . .', suggesting, perhaps, that if it is metaphysically possible to extract and restore brain information, then it is also metaphysically possible for someone to be given a new brain. Williams might be interpreted here as denying even the essentiality of a portion of one's brain. (As a Reductionist, however, Williams would be forced to regard personal identity in such a case as consisting in something else—hence the remark about the identity of the rest of the body as something to 'hold onto.')

John Perry suggests a principle similar to the essentiality of one's brain and proceeds to criticize it, in his *A Dialogue on Personal Identity and Immortality* (Indianapolis: Hackett, 1978), at p. 47. Perry is also a Reductionist regarding persons (see, for example, pp. 21–22 of the same work), and his criticisms suggest that his intended target is a principle supporting a brain-based criterion of personal identity as a solution to the Reductionist problem.

[33] The combined principle of the essentiality of brain ownership is similar to what Parfit, in *Reasons and Persons*, p. 204, calls *the Physical Criterion* (although the latter is actually a principle of the essentiality of unique ownership of some sufficient portion or other of a brain). The thesis I am proposing is significantly different from the view of Thomas Nagel in *The View From Nowhere* (Oxford University Press, 1986). Nagel is at least tempted to identify a person with his/her brain, while denying that the connection is *a priori*. On his view, the sentence 'Jones = Jones's brain', and likewise the sentence 'Jones = *B*' where '*B*' names Jones's brain, express necessary *a posteriori* truths. I am claiming that though a person and his/her brain are not identical, they are essentially related to each other by functional ownership. On my view, the sentence 'Jones's brain = *B*'—or more cautiously, the sentence 'If Jones has a functioning brain, then it is a portion of *B*'—expresses a necessary *a posteriori* truth, whereas Nagel's allegedly necessary *a posteriori* sentences are not even true.

On the other hand, Nagel describes the identification of person and brain as a 'mild exaggeration' (p. 40). Nagel's actual view may thus be closer to the view defended here. Parfit reports (*ibid.*, pp. 289–293, 468–477), that in then unpublished work (possibly a draft of chapter 3 of *The View From ow here*), Nagel rejects the Extreme Haecceitist thesis that the identity of a person, given qualitatively, is a further fact about the person given thus. Nagel reportedly defends a brain-*qua*-organ-based criterion as a solution to the meta-question for a Reductionist problem of personal identity. It is possible, however, that Nagel endorses a Mereological Reductionism of the sort described in note 18 above. Being an Extreme Haecceitism, I reject the traditional Reductionist problem of personal identity as a pseudo-problem. I similarly reject the idea that the identity of the person having a particular brain is mereologically reducible to facts about the brain. Whatever force Parfit's objections may have against Nagel's reported view, they carry little or no weight against my proposed essentialist principles. Another account having important points of contact with the account presented here is that in Peter Unger, *Identity, Consciousness, and Value* (Oxford University Press, 1990)—with the significant difference that Unger (p. 42) declines to endorse any nontrivial essentialism of the sort that is central to the present view.

VIII

The twin principles of the essentiality of one's brain and the essentiality of a brain's ownership yield answers to the metaphysical meta-question on the Essentialist version of the problem of personal identity. They also yield answers to the primary questions about D/R-1 and D/R-2, interpreted on the Essentialist scheme, different answers to each. Scenario D/R-3 is another matter. The essentialist principles seem not to help at all in settling the primary question about D/R-3—even when interpreting it as a question about the genuinely possible worlds represented by the scenario. We are not told whether the changes that have taken place in D/R-3 are the result of brain transplants, or information extraction by means of a BW device, or alien rays, or magic (if such is possible), or something else. The body-based and the psychology-based criteria go about their business in D/R-3 just as they do in D/R-1 and D/R-2, making the same identifications in all three scenarios. But the brain-*qua*-organ criterion comes up short in D/R-3. The criterion is not up to the task of answering the primary question about the most puzzling of puzzle cases, interpreted on the Essentialist scheme. There seems to be nothing in this neutral scenario for the criterion to take hold of.

There is something. The brain-*qua*-organ criterion discriminates between D/R-1 and D/R-2, making opposite identifications in each. That in itself, I have argued, is an important feature of the criterion. And it is a feature that the criterion brings with it to D/R-3. D/R-3 is neutral regarding the sort of facts in virtue of which D/R-1 and D/R-2 differ from each other. The reason the brain-*qua*-organ criterion is unable to settle the primary question in D/R-3 is that D/R-3 is silent where D/R-1 and D/R-2 are specific. D/R-3 fails to specify the sort of facts that the brain-*qua*-organ criterion needs in order to identify the post-switch body-persons. If D/R-3 is brought about by brain transplants performed by alien surgeons, then the identifications go the same way as in D/R-1. If D/R-3 is brought about instead through an alien version of the BW device, then the identifications go the same way as in D/R-2. Some facts or other of this sort must obtain in any scenario that realizes D/R-3. Yet D/R-3 fails to specify what they are.

The difficulty encountered in the attempt to answer the primary question creates the impression that one is confronting a deep philosophical conundrum for which the brain-*qua*-organ criterion's effectiveness breaks down and is seen to be inadequate. But the difficulty (indeed impossibility) of answering the primary question about D/R-3 is not due to a defect in the brain-*qua*-organ criterion. It is due to a defect in the scenario. It is under-specified, and for that very reason the brain-*qua*-organ criterion yields no answer to the primary question. In truth, D/R-3 is not so much a particular puzzle-case scenario for the Essentialist problem of personal identity as it is a generic *category* or *classification* of puzzle cases. It represents a *class* or *type* of genuinely possible scenarios. In some possible scenarios of that type, Doris and Rita exchange bodies while retaining their brains, and consequently also their psychologies. In other possible scenarios *of the very same type*, Doris and Rita exchange their psychologies while retaining their bodies, including their brains.

In some possible D/R-3 scenarios the Doris-body-person is Doris; in other possible D/R-3 scenarios the Doris-body-person is Rita. The reason the best criterion yet considered does not yield a single, unequivocal answer to the primary question, interpreted on the Essentialist scheme, when asked of D/R-3 itself, is that no answer is correct for all such cases. Any criterion that provides a single answer for all genuinely possible puzzle cases of that type is *ipso facto* mistaken. This would include both the body-based and the entire array of psychology-based criteria.[34]

This is true to a lesser extent about D/R-1 and D/R-2 as well. D/R-1 and D/R-2 are also under-specified, indeed in infinitely many respects. D/R-1 fails to specify, for example, the details of the surgical procedure that Allen performs on his victims. And D/R-2 does not specify very much at all about how the marvelous BW device works. The Doris/Rita incidents are not so much particular puzzle-case scenarios for the Essentialist problem of personal identity as they are *types* of puzzle cases. But there is an important difference between D/R-1 and D/R-2 on the one hand, and D/R-3 on the other. Arguably, all possible D/R-1 cases yield the same identifications, and all possible D/R-2 cases also yield the same identifications, exactly opposite to those of D/R-1 cases. The class of genuinely possible worlds represented by D/R-1 is a uniform class in regard to the relevant identifications. Similarly for the class of genuinely possible worlds represented by D/R-2. D/R-3, by contrast, defines a remarkably mixed bag. D/R-3 fails to specify the very sorts of facts upon which the answer to the primary question supervenes—to wit, the matter of whether the Doris-body-person's consciousness is being produced by what had been Doris's brain or by what had been Rita's. The class of genuinely possible worlds represented by D/R-3 remains diverse, in the same way as the class of all worlds, possible and impossible, represented by either D/R-1 or D/R-2. The primary question about D/R-3, even when interpreted on the Essentialist scheme, is thus a 'wife-beating' question. It remains unanswerable because of its false presupposition that the answer modally supervenes on facts concerning psychologies and/or bodies.

[34] For related discussion, see *Reference and Essence*, at pp. 242–246.