The Philosophical Quarterly, *Vol. 52, No. 206* *January 2002*
ISSN 0031–8094

PERSONAL IDENTITY AND THOUGHT-EXPERIMENTS

By Tamar Szabó Gendler

Through careful analysis of a specific example, Parfit's 'fission argument' for the unimportance of personal identity, I argue that our judgements concerning imaginary scenarios are likely to be unreliable when the scenarios involve disruptions of certain contingent correlations. Parfit's argument depends on our hypothesizing away a number of facts which play a central role in our understanding and employment of the very concept under investigation; as a result, it fails to establish what Parfit claims, namely, that identity is not what matters. I argue that Parfit's conclusion can be blocked without denying that he has presented an imaginary case where prudential concern would be rational in the absence of identity. My analysis depends on the recognition that the features that explain or justify a relation may be distinct from the features that underpin it as necessary conditions.

It is good to know something of the customs of various peoples, in order to judge our own more objectively, and so that we do not make the mistake of the untravelled in supposing that everything contrary to our customs is ridiculous and irrational. But when one spends too much time travelling, one becomes at last a stranger at home.... In addition, fiction makes us imagine a number of events as possible which are really impossible.... Thus it happens that those who regulate their behaviour by the examples they find in books are apt to fall into the extravagances of the knights of romances, and undertake projects which it is beyond their ability to complete (Descartes, *Discourse on the Method*, I 6–7).

I. INTRODUCTION

As things stand, fission and fusion and teletransportation are the stuff of science fiction. Non-fictional infants are produced by a well known sequence of biological processes, and non-fictional adults develop from non-fictional infants by an equally well known sequence of biological and social processes. Moreover, each non-fictional adult lives life knowing that this is how things are: one of the facts of life is that (like storks, but unlike *in vitro* fertilization) fission and fusion and teletransportation do not play any role in bringing actual persons into being.

Recent philosophical discussions of the nature and value of personal identity, however, have tended to treat these 'facts of life' as *provincial* truths

– as facts about persons-as-they-happen-to-be, not facts about persons-as-they-really-generally-are.[1] And in an effort to avoid making 'the mistake of the untravelled', that is, the mistake of taking local customs to be universal practices, appeal is made to imaginary cases, cases which, like experiments, are supposed to help to compensate for the often arbitrary ways in which we come upon information in the world.

My goal in this article is to explore how this methodology may fall foul of the following principle: where contingent correlations play a central role in our understanding and employment of a concept, and where the concept in question concerns an assessment of value, cases where we 'imagine away' the correlations may be uninformative as guides to that concept's application-conditions. The fact that two features coincide in all actual cases may mean that there is no straightforward way for us to determine how we would or should respond to either in isolation.

I shall explore this general principle by looking at a specific case, Derek Parfit's widely discussed argument concerning the importance of personal identity. What I shall argue is that Parfit's argument is unsuccessful because it depends on ignoring the provincial truths adverted to in the opening paragraph – truths describing the contingent concomitance of a cluster of features that can, in imaginary cases, be conceptually separated. In his discussions of personal identity, Parfit seeks to establish that what ought rationally to matter to us when we consider survival and future well-being is not that *we ourselves* should survive, but only that people should exist who are psychologically continuous with us in the right sort of way. His methods for doing so involve describing imaginary cases where it seems clear that identity would not be what matters in this way, and arguing that if we are consistent in our commitments, then we ought to conclude that even in actual cases identity is not what matters.[2]

If I am right, however, Parfit's conclusion can be blocked without denying that he has presented an imaginary case where prudential concern would be rational in the absence of identity. (Throughout this article I follow the recent personal identity literature in using 'prudential concern' as a term of art, referring to the sort of concern that we bear towards our future selves. In so doing, I do not mean to deny that we might in the ordinary sense of the term have prudential concern for others, such as our children or our

[1] See, e.g., Shoemaker in S. Shoemaker and R. Swinburne, *Personal Identity: Great Debates in Philosophy* (Oxford: Blackwell, 1984), p. 127; P. Unger, *Identity, Consciousness and Value* (Oxford UP, 1990), p. 11; R. Nozick, *Philosophical Explanations* (Cambridge: Belknap, 1981), p. 30.

[2] Parfit repeatedly speaks of his conclusion as being that '*personal identity is not what matters*' (his italics): see, for instance, Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984), p. 255. Though I shall argue below that the expression 'what matters' is ambiguous in certain ways, for the time being I take it as a place-holder.

friends.[3]) If, as I shall argue, the feature that explains or justifies a relation can be distinct from those features that underpin it as necessary conditions, then we can accept that there could be cases of the sort Parfit describes, without taking them to have the implications which he takes them to have.

In the first main section (§II), I shall present Parfit's argument, and identify three crucial assumptions that underlie his reasoning:

Intrinsicness principle: the relation that matters for rational prudential concern, *M*, is an intrinsic relation[4]

Sufficiency principle: if the relation that matters for rational prudential concern (*M*) holds between *A* and *B*, then *A*'s prudential concern for *B* is rational

Necessity principle: if *A*'s prudential concern for *B* is rational, then the relation that matters for rational prudential concern (*M*) holds between *A* and *B*.

Together, these principles generate a particular internalist commitment: they imply that there is a certain relation *M* which is both necessary and sufficient for rational prudential concern (of *A* for *B*), a relation whose obtaining depends only on facts about *A* and *B* and the resulting relations between them. What I try to show is that this commitment is reasonable on one understanding of what *M* involves, but not on another, and that Parfit's argument rests on an equivocation between the two senses.

The first sense of *M* in which we might be interested is this: we might be interested in finding out which, if any, relation is common to all (possible) cases where rational prudential concern obtains. I shall call this the 'common factor' sense. But there is also a second sense of *M* in which we might be interested: we might be interested in finding the relation that *explains* why rational prudential concern obtains. I shall call this the 'explanatory' sense. Of the three principles just enumerated (the intrinsicness, sufficiency and necessity principles), all three are true when we are interested in *M* in the common factor sense.[5] But if we are interested in *M* in the second,

[3] See also M. Schechtman, 'Personhood and Personal Identity', *Journal of Philosophy*, 77 (1990), pp. 71–92, and *The Constitution of Selves* (Cornell UP, 1996); J.D. Velleman, 'Self to Self', *Philosophical Review*, 105 (1996), pp. 39–76; J. Whiting, 'Friends and Future Selves', *Philosophical Review*, 95 (1986), pp. 547–80; S. Wolf, 'Self-Interest and Interest in Selves', *Ethics*, 96 (1986), pp. 704–20.

[4] Like others in the personal identity literature, I rely here on an informal notion of intrinsicness: cf. Wiggins' *only a and b* principle, in D. Wiggins, *Sameness and Substance* (Harvard UP, 1980), p. 96; Noonan's *only x and y* principle, in H. Noonan, *Personal Identity* (London: Routledge, 1989), p. 16; Nozick's relevance principle, in his *Philosophical Explanations*, p. 31; and Parfit's intrinsicness of personal identity principle, *Reasons and Persons*, p. 267.

[5] In the interests of simplicity, I am assuming that *M* is believed to obtain iff *M* obtains (this bears on the truth of sufficiency). I am also assuming that disjunctions of intrinsic relations are themselves intrinsic (this bears on the truth of the intrinsicness principle if we allow that *M* may be disjunctive).

explanatory, sense, neither the intrinsicness principle nor the necessity principle is true. Part of the apparent success of Parfit's argument can be attributed to a failure to make this distinction. I address this in §II.

But while making this distinction allows us to see a certain conceptual possibility, it does not seem to capture much about why the argument *feels* so convincing. My aim in the remainder of the article is to diagnose the source of its apparent persuasiveness, and to show that this appearance is misleading. I shall do so by suggesting that the reasoning on which Parfit depends rests on a seemingly undeniable principle of rationality, what Mill called 'the method of agreement'. The method of agreement says roughly that if there is a single feature whose presence or absence directly correlates with the presence or absence of the phenomenon under scrutiny, then it is this feature that underlies the obtaining of the phenomenon. Without denying the general legitimacy of this form of reasoning, I suggest that it is misapplied in Parfit's case. For when we are concerned with grounds for explanation (as opposed to necessary and sufficient conditions), and where one of the competing explanations is a special case of the other, Mill's test is irredeemably inconclusive. But, I shall go on to argue, there is a second test which we can use to decide between the two explanations, a test I call 'the association-dependence test', which tells in favour of my interpretation rather than Parfit's. I present my negative argument against Parfit in §§III–IV, and the positive argument for my own view in §V.

II. THE ARGUMENT AND ITS CRUCIAL ASSUMPTIONS

II.1. *Parfit's fission argument*

Parfit's familiar fission argument can be reconstructed as follows.[6] Triplets are involved in an accident, in which the body of one, Brainy, is fatally injured, while the brains of his two brothers are totally destroyed. Brainy is such that the physical bases for his psychological characteristics are realized in duplicate, one complete set in each lobe. Following the accident, doctors divide his brain in half, and transplant the two hemispheres into the bodies of the two brothers.

In the first scenario, which I shall call 'the single transfer case', only the left transplant succeeds, and the right transplant is destroyed. The resulting individual, Lefty (this term is an abbreviation for 'the individual who has Brainy's original left hemisphere'), has all of Brainy's memories and psychological states and a body almost indistinguishable from the one that Brainy

[6] See Parfit's canonical presentation of the case at *Reasons and Persons*, pp. 254–5. I am granting for the sake of argument that the scenario described is coherent.

had before the accident. Parfit holds, and for the sake of argument I shall grant it to him, that

1. In the single transfer case, Lefty *is* Brainy.[7]

Given this, along with the principle that

2. If *A* is identical with *B*, then the relation that matters for rational prudential concern, *M*, holds between *A* and *B*

it follows trivially that

3. In the single transfer case, *M* holds between Brainy and Lefty.

This process of reasoning is intended to establish a base case for an argument from parity. The parity argument concerns a second scenario, 'the double transfer case'. Here both transplants are successful. Each of the two resulting individuals, Lefty and Righty, has all of Brainy's memories and psychological states and a body almost indistinguishable from the one that Brainy had before the accident. Parfit points out that

4. Brainy's relation to Lefty is intrinsically the same in the single and double transfer cases.

From (3) and (4), together with

5. *The intrinsicness principle* (for *M*): *M* is an intrinsic relation

there follows

6. *The parity result* (for *M*): if *M* holds between Brainy and Lefty in the single transfer case, then *M* also holds between Brainy and Lefty in the double transfer case.

And from (3) and (6) it follows that

7. *M* holds between Brainy and Lefty in the double transfer case.

I shall assume that Brainy is a single person,[8] and also that Lefty and Righty are not the same person. After all, the two occupy distinct spatial

[7] For representative challenges, cf. B. Williams, *Problems of the Self* (Cambridge UP, 1973), essays 1–5; J.J. Thomson, 'People and Their Bodies', in J. Dancy (ed.), *Reading Parfit* (Oxford: Blackwell, 1997), pp. 202–29; E. Olson, *The Human Animal: Personal Identity without Psychology* (Oxford UP, 1997); and many of the contributions to D. Cockburn (ed.), *Human Beings* (Cambridge UP, 1991).

[8] For challenges, see D. Lewis, 'Survival and Identity' (1976), repr. in his *Philosophical Papers*, Vol. 1 (Oxford UP, 1983), pp. 55–72, with postscript at pp. 73–7; Noonan, *Personal Identity*, pp. 164–8, 197–8; E. Mills, 'Dividing without Reducing: Bodily Fission and Personal Identity', *Mind*, 102 (1993), pp. 37–51; D. Robinson, 'Can an Amoeba Divide without Multiplying?', *Australasian Journal of Philosophy*, 63 (1985), pp. 299–319.

locations, undergo different experiences, and have no unusual causal effect on one another. If Lefty and Righty are different people, and Brainy is a single person, Lefty and Righty cannot both be identical with Brainy. So

8. In the double transfer case, Lefty *is not* Brainy.

(Strictly speaking, of course, additional assumptions are required for this step; I here omit the reasoning in the interests of space.) But since (7) tells us that the relation that matters for rational prudential concern holds between them, it follows that

9. In the double transfer case, *M* is not identity.

But *M* is univocal, so

10. In the single transfer case, *M* is not identity

and, more generally,

11. *The unimportance of identity conclusion*: *M* is not identity.

II.2. *Three crucial distinctions*

Despite its apparent clarity, Parfit's argument persuades only by blurring three crucial distinctions, which I discuss in turn in this section. The first is between two sorts of states of affairs, those where whatever relation it is that 'matters' for prudential concern holds between *A* and *B* (I shall have more to say about what this means at the end of this subsection), and those where *A*'s prudential concern for *B* is rational, in the sense that the norms of rationality permit *A* to be prudentially concerned for *B*.[9] In shorthand, I shall speak of

(a) *M*(*A*,*B*): the relation that matters for rational prudential concern holding between *A* and *B*
(b) *RPC*(*A*,*B*): *A*'s prudential concern for *B* being rational.

It has generally been assumed that these two relations coincide – that wherever there is *RPC* (that is, in any case where we are warranted in judging that prudential concern by *A* for *B* is rational), there must be *M* (that is, the relation that matters for prudential concern must obtain between *A* and *B*), and correspondingly wherever there is *M* there must be *RPC*. My

[9] The norms of rationality *permitting A* to be prudentially concerned for *B* are *prima facie* different from the norms of rationality *requiring A* to be prudentially concerned for *B*. (I thank Carol Rovane for pressing me on the need to make this distinction.) I suspect that the relation between prudential concern and rationality is such that if prudential concern is rationally permitted then it is rationally required, but my argument below is non-committal on this question.

second crucial distinction expresses these dependencies:

(c) *Necessity principle*: if *A*'s prudential concern for *B* is rational, then the relation that matters for rational prudential concern holds between *A* and *B*: i.e., $RPC(A,B) \rightarrow M(A,B)$

(d) *Sufficiency principle*: if the relation that matters for rational prudential concern holds between *A* and *B*, then *A*'s prudential concern for *B* is rational: i.e., $M(A,B) \rightarrow RPC(A,B)$.

For the remainder of my discussion, I shall assume the truth of the sufficiency principle. But to say whether I think the necessity principle is true, I need to make one final distinction.

'The relation that matters for rational prudential concern' (granting hyperintensionalism) might be taken in at least two ways:

(e) *Common factor reading* (for *M*): *the relation that matters for rational prudential concern* is some relation that is common to all and only cases where rational prudential concern obtains

(f) *Explanatory reading* (for *M*): *the relation that matters for rational prudential concern* is some relation that explains the rationality of prudential concern obtaining in cases where rational prudential concern obtains (for a rough characterization of 'explains', see § III and IV below).

The distinction between these two readings bears some relation to Aristotle's distinction between efficient and formal causes.

On the common factor reading (e), the necessity principle (c) is trivially true. If 'the relation that matters for rational prudential concern' is some relation common to all and only cases where rational prudential concern obtains, then clearly *A*'s prudential concern for *B* is rational (*RPC*) only if the relation that matters for rational prudential concern (*M*) holds between *A* and *B*. But on the explanatory reading (f), the connection is not so clear. If 'the relation that matters for rational prudential concern' is some relation that explains the rationality of prudential concern obtaining in cases where rational prudential concern obtains, it does not follow (so I shall argue in §V) that *A*'s prudential concern for *B* is rational only if *M* holds between them.

II.3. *The intrinsicness premise*

With these distinctions in place, I can return to Parfit's argument. The crucial premise in Parfit's argument is, of course,

5. *The intrinsicness principle* (for *M*): *M* is an intrinsic relation.

It is the intrinsicness principle (for *M*) that allows Parfit to derive (6) the parity result (for *M*), and hence (11) the unimportance of identity conclusion.

Parfit's argument goes through if *M* is understood according to the common factor reading; on that reading the intrinsicness principle is true, the parity result follows, and the unimportance of identity conclusion is thereby established. But for Parfit's purposes the common factor reading is inadequate. On this reading, the unimportance of identity conclusion says merely that identity is not common to all cases where there is rational prudential concern. But Parfit needs more than this: if we respond to the case in the way he expects, he thinks we should change our views about what underpins our prudential concern for our future selves. To show this, he needs to show that identity does not explain the rationality of prudential concern. But when *M* is understood in this sense, I shall contend, the intrinsicness principle is false. (I shall argue for this in the sections that follow.) And if this is so, then Parfit's argument has not succeeded in showing what it has been taken to show.

II.4. *Summary*

Parfit's goal is to show that 'personal identity is not what matters' or, in my terms, that the relation that matters for rational prudential concern is not identity. He seeks to do this by presenting a case (the double transfer case) where it seems that what matters for rational prudential concern is present, while identity is absent. And if the former can obtain without the latter, then identity cannot be what matters for rational prudential concern. I have suggested a number of distinctions that need to be drawn in evaluating the force of this argument. I have pointed out that we need to separate questions about the relation I have been calling *RPC* (the relation that holds when *A*'s prudential concern for *B* is rational) from questions about the relation which I have been calling *M* (the relation that matters for rational prudential concern). And I have pointed out that *M* can be understood on either a common factor or an explanatory reading, and claimed that if Parfit wishes to establish the conclusion he does, then we need to understand *M* according to the explanatory reading. With these distinctions in place, I have suggested, we can diagnose part of the force of Parfit's argument. Fission does describe a case in which Brainy bears a relation of rational prudential concern (*RPC*) for a continuer with whom he is not identical. And if *RPC* is necessary and sufficient for *M*, then if we grant (as I am willing to do) that the intrinsicness principle is true for *RPC*, Parfit's argument goes through. But, I shall argue, *RPC* is necessary and sufficient for *M* only on the common factor reading, not on the explanatory reading. I hypothesize that the argument's apparent undeniability comes from a failure to make these distinctions. It is because we fail to see that *RPC* and *M* could come apart in the way I am suggesting that Parfit's conclusion seems mandatory.

III. WHY IS THE FISSION ARGUMENT SO COMPELLING?

III.1. *Diagnosis*

But what explains our failure to see this? Why are we so convinced that *M* and *RPC* must co-vary, even once we have distinguished the common factor and explanatory readings? Why does the necessity principle seem so hard to reject? I think the answer can be traced to a tendency to misapply an otherwise legitimate principle of scientific reasoning which Mill called the *method of agreement*. Mill's principle says 'If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause (or effect) of the given phenomenon'.[10] The principle is often generalized to cover phenomena other than cause and effect, and it is a cousin of Mill's principle which concerns us here. Following Mill, this principle might be put as follows: 'If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the *explanation* for the given phenomenon'. It is this principle to which I think we tacitly appeal when we are convinced that *M* and *RPC* must co-vary. But this principle, as I shall argue below, cannot be straightforwardly applied to Parfit's case.

III.2. *The appeal of the principle*

It is worth spending a moment getting clear about why Mill's principle seems so central to scientific reasoning, starting with causes. Suppose that whenever I strike a match against the side of a matchbox and say 'Let there be light', the match bursts into flame; whenever I strike a match against the side of a matchbox and say nothing, the match bursts into flame; whenever I simply hold the match in the air and say 'Let there be light', the match remains unlit; and whenever I neither strike the match nor recite the incantation, the match remains unlit. That is:

Match-and-incantation case

A = strike match against box
B = utter 'Let there be light'
P = match bursts into flame

	A	not-*A*
B	*P*	not-*P*
not-*B*	*P*	not-*P*

[10] J.S. Mill, *A System of Logic*, III viii 2, in his *Collected Works*, Vol. VII, ed. J.M. Robson (Univ. of Toronto Press, 1973), p. 390.

A quick glance at the chart reveals that *P* obtains when and only when *A* obtains, and that the obtaining of *P* is indifferent to the obtaining of *B*. Assuming certain conditions – that *A* and *B* are the only relevant factors in this case; that *A* brings about *P* in the same way in the *A*-alone case and in the *A*-plus-*B* case; that *A*, *B* and *P* are specified at commensurate levels of description, etc. – this pattern gives us reason to conclude that *A* explains the obtaining of *P*, and that *B* does not. Unless we are inclined to serious scepticism, it seems irrational to insist that it is the incantation, or even the incantation-plus-striking, that explains the match's bursting into flame.

III.3. *Fission and the method of agreement*

I am now in a position to state more precisely what goes awry in the Parfit case, namely, that those who find themselves compelled by his argument mistakenly assimilate it to something like the match-and-incantation case. So I shall present the case as I think they see it, and then explain why I think the analogy fails.

The discussion here will be helped by the introduction of a few terms of art. Parfit holds that the relation that obtains between *X* and *Y* in all (real and imaginary) cases where *X*'s prudential concern for *Y* is rationally warranted (the 'common factor' of my discussion above) is something he calls *relation R*. (My argument is indifferent to the particular features of this relation, but it may be helpful to note that Parfit at p. 215 characterizes relation *R* as 'psychological continuity with the right sort of cause'.) Let us grant, if only for the sake of argument, that personal identity over time can be analysed as the non-branching holding of relation *R*: *Y* is identical to *X* iff *X* is *R*-related to *Y*, and *X* is not *R*-related to any non-*Y* that co-exists with *Y*. (I am skating over delicate metaphysical issues here in the interests of making a more general point about methodology.) I conceded in §II that whenever both relation *R* and absence-of-competitors obtain (that is, wherever *X* and *Y* are identical), rational prudential concern is warranted. And I also conceded there that in the double transfer case, where relation *R* obtains *without* the absence of a competitor, rational prudential concern may well be warranted. Moreover, I shall also concede that rational prudential concern is never warranted in the absence of *R*. That is:

Fission case

A = relation *R* obtains between *X* and *Y*
B = relation *R* does not obtain between *X* and any non-*Y* that co-exists with *Y*
P = *X*'s prudential concern for *Y* is rational

	A	not-*A*
B	*P*	not-*P*
not-*B*	*P*	not-*P*

As in the match-and-incantation case, *P* obtains in the fission case when and only when *A* obtains; and as in the match-and-incantation case, the obtaining of *P* is indifferent to the obtaining of *B*, in the sense that *B*'s presence is neither necessary nor sufficient for the obtaining of *P*. As before, assuming (for the time being) certain conditions – that *A* and *B* are the only relevant factors in this case; that *A* brings about *P* in the same way in the *A*-alone case and in the *A*-plus-*B* case; that *A*, *B* and *P* are specified at commensurate levels of description, etc. – this pattern seems to give us reason to conclude that it is *A*, not *B*, and not even *A*-and-*B*, that explains the obtaining of *P*. To quote Parfit (in an unpublished paper): 'In all ordinary cases, personal identity and [*R*] coincide. When they diverge, [*R*] is what matters. That strongly suggests that, in all cases, [*R*] is what matters.... If, when two facts come apart, one of them is what matters, why think the *other* is what matters when they coincide?'

III.4. *Disanalogies*

There are at least three specific ways in which the fission case differs from the match-and-incantation case and which are relevant to whether the method of agreement can be properly employed: these concern *subject-matter*, *internal structure*, and *background conditions*. (There is a fourth, more general, worry which I have bracketed in the context of this discussion, the worry that the sort of factorization presupposed by this account is even possible in the case of the concept of personal identity.)

(a) *Subject-matter*: in the match-and-incantation case we are concerned with *causal* explanation, whereas in the fission case we are concerned with the explanation of *value*

(b) *Internal structure*: in the match-and-incantation case there are two independent features at play whose relative contributions we are trying to deduce, whereas in the fission case the competing explanations are *A* and *A*-plus-*B*, where *A*-plus-*B* is a *specification* of *A* (being a unique *R*-related continuer is a way of being an *R*-related continuer)

(c) *Background conditions*: in the match-and-incantation case, the factors under scrutiny frequently arise independently, whereas in the fission case we are concerned with a situation where in all actual instances the factors we are considering coincide.

What I shall argue in §IV (employing the very method under discussion!) is that it is the conjunction of (a) and (b) that makes the methodology indecisive in the Parfit case. I shall then go on in §V to discuss the relation of this to (c).

IV. WHERE DOES THE ANALOGY GO AWRY?

IV.1. *Explanatory subject-matter*

I note first that the method of agreement is perfectly legitimate in cases of evaluative explanation, so long as the factors being disambiguated by its means are genuinely independent. That is, in cases that resemble the fission case as far as (a) is concerned, but resemble the match-and-incantation case as far as (b) and (c) are concerned, there is no problem with using the methodology. Suppose, for instance, we are trying to determine why I love 'The Star-Spangled Banner' so much – is it the stirring tune, or the inspiring words? And suppose we conduct a test with the following results:

Star-Spangled Banner case

A = *X* has the same words as 'The Star-Spangled Banner'
B = *X* has the same tune as 'The Star-Spangled Banner'
P = *X* fills me with admiration, awe and joy

	A	not-*A*
B	*P*	not-*P*
not-*B*	*P*	not-*P*

For reasons parallel to those adduced above (whenever I hear the words I am filled with awe and admiration, regardless of the tune to which they are sung, and whether the tune inspires this sentiment depends on whether it is sung with those words), it seems reasonable to conclude that what explains my admiration for 'The Star-Spangled Banner' are Francis Scott Key's magnificent verses, and not the tune of John Stafford Smith's rousing 'To Anacreon in Heaven'. Assuming as before that *A* and *B* are the only relevant factors in this case, that *A* brings about *P* in the same way in the *A*-alone case and in the *A*-plus-*B* case, and that *A*, *B* and *P* are specified at commensurate levels of description, there is no problem as such with applying the method of agreement to cases where what is at issue is the explanation of value.

IV.2. *Genus and species*

Though certain complications arise, there is also no reason to doubt the general validity of the methodology when we are considering cases that resemble the match-and-incantation case as far as (a) and (c) are concerned, but are like the Parfit case in terms of (b), that is, cases where the competing explanations are not *A* and *B*, but rather *A* and *A*-plus-*B*, and where *A* and *A*-plus-*B* are related as genus and species:

Boat-sinking case 1

A = X weighs at least 20 pounds
B = X does not have a weight other than 20 pounds
P = placing X on the boat causes the boat to sink

	A	not-A
B	P	not-P
not-B	P	not-P

Given the circumstances set out in the table, it seems *ceteris paribus* reasonable to conclude that what causes the boat to sink is an object's weighing at least 20 pounds.

Although it is technically satisfactory, this way of setting up the case clearly feels contrived: where the A-plus-B case is a special instance of the A case (weighing exactly 20 pounds is a special way of weighing at least 20 pounds), shoe-horning it into a fourfold matrix means specifying the B condition in a way that allows it to be vacuously satisfied whenever A does not hold. What this means is that the information we obtain distinguishes among what seem to be three rather than four possibilities. So we do better to represent the case as follows:

Boat-sinking case 2

A = X weighs at least 20 pounds
C = X weighs exactly 20 pounds
P = placing X on the boat causes the boat to sink

	A	not-A
C	P	
not-C	P	not-P

The non-independence of A and C means that this test is less informative than the four-way test: because there is no case where we have C without A, there is no information to be had about what C alone contributes. But in the boat-sinking case, and, I hypothesize, in cases of (non-reason-involving) causal explanation in general, this does not undermine the method's reliability. The structure of causal explanation gives us antecedent reason to assume that the way A brings about P in the A-plus-C case is (so far as sinking is concerned) the same as the way A brings about P in the A-not-C case. Causal explanation presupposes that the general supersedes the particular, and in such cases the method of agreement provides us with reliable grounds for taking one or another feature as explanatory.

IV.3. *'Borrowed lustre'*

By contrast, reason-explanation makes no such presupposition, and as a result cases where value-explanatory subject-matter is combined with genus–species inner structure are ill suited to investigation by the method of

agreement. For the method is ineffective in cases involving what I shall call 'borrowed lustre', where both pure and impure instances of a phenomenon are accorded the same assessment because impure instances are treated as relevantly similar to pure ones.

An example may make this clearer. Suppose we venerate regular geometrical figures for their beauty, but certain approximations to regular figures also produce the same respect by way of resembling the ideal. We might portray the circumstances as follows:

Square case

A = X is square-like
C = X is square[11]
P = X is an appropriate target of geometrical veneration

	A	not-A
C	P	
not-C	P	not-P

What the chart reveals is that whenever something is square-like, it is an appropriate target of geometrical veneration, and whenever something is not square-like, it is not an appropriate target of geometrical veneration; whether or not it is actually square has no bearing on its suitability as an object of geometrical veneration. But from this we are not entitled to conclude that it is square-likeness rather than squareness that explains the appropriateness of geometrical veneration; as I stipulated in the formulation of the case, what explains appropriate geometrical veneration is regularity – and approximate squares come to be appropriate objects of geometrical veneration by way of resemblance to the ideal.

The reason for this is that in borrowed-lustre cases one of the antecedent conditions for application of the method is not satisfied: the way in which *A* brings about *P* in the *A*-plus-*C* case is different from the way *A* brings about *P* in the *A*-not-*C* case. When I venerate a perfect square, its square-like features (*A*) cause me to venerate it (*P*) because of their resemblance to a feature that it has, squareness; whereas when I venerate a merely approximate square, its square-like features (*A*) cause me to venerate it (*P*) because of their resemblance to a feature that it lacks, squareness.

Lest you think this case is anomalous, I offer three additional cases, which different readers appear to find persuasive to different degrees.[12] The cases

[11] We could represent this as a four-way matrix if we define *B* as '*X* lacks all non-square geometric features'. In this case the matrix will be filled out as in the match-and-incantation, fission and boat-sinking 1 cases. As in the boat-sinking 1 case, the *B*-not-*A* case will be satisfied only by things to which *A* fails to apply: whatever lacks all non-square geometric features without being square-like will lack all square-like features whatsoever.

[12] One might think of these as being cases involving what Robert Nozick has called 'symbolic utility': Nozick, *The Nature of Rationality* (Princeton UP, 1993), esp. pp. 26–35.

are presented somewhat sketchily: their purpose is just to make clear the range of instances where the phenomenon in question seems to arise.[13]

Dead body case

A = X is a human body
C = X is a living human body
P = X is an appropriate object of respect (*qua* body)

Stuffed animal case

A = X has the appearance of a baby seal
C = X is a baby seal
P = X is not something that should be hurled across the room

	A	not-A
C	P	
not-C	P	not-P

Vegetarianism case

A = X resembles a piece of meat from a cow subjected to conditions of factory farming
C = X is a piece of meat from a cow subjected to conditions of factory-farming
P = X is something that should not be eaten by me

I do not deny that objections can be raised for each of these particular cases: that the degree of respect required (*qua* body) may be higher in the case of a living body than a dead one; that things in general should not be hurled across the room, so *a fortiori* stuffed seals should not be; and so on. In each case, I suspect that appropriate modifications could be made so as to retain the larger point, but this would be at the expense of presenting the cases schematically.

At any rate, resistance to the details of one or another particular case should not get in the way of understanding what they are intended to illustrate. The principle that lies behind borrowed-lustre cases is that when we are concerned with the explanation of value, it is possible to 'cantilever' out from central or ideal cases (perfect squares, living bodies, real baby seals, factory-farmed meat) to peripheral cases (approximate squares, dead bodies, stuffed baby seals, non-factory-farmed meat) in a way to which the method of agreement is indifferent. In short, the method cannot tell us whether it is

[13] As before, these cases can be represented on a four-way matrix with *B*-conditions roughly like the following (in order to provide clean 'No' answers to the *B*-not-*A* case, corresponding modifications will need to be made to the characterization of *P*): dead body case: B = 'X is living'; stuffed animal case: B = 'X has an internal structure appropriate to the kind that it appears to be'; vegetarianism case: B = 'X is an instance of what it resembles'.

the special case that explains the general one, or *vice versa*; and, I have argued, there are cases where the former rather than the latter holds.

V. THE POSITIVE ARGUMENT

V.1. *The association-dependence test*

As far as the method of agreement is concerned, then, we seem to be at an *impasse*; Parfit's argument is inconclusive, and I have shown why. But thus far I have offered no reason to think that relation *R*'s underpinning of prudential concern is the result of borrowed lustre. And without such an argument I can do no better than parity.

I suggest that there is a second test that can help us to distinguish between cases where it is *A* (the general case) that explains our evaluation of *C* (the special case), and cases where it is *C* (the special case) that explains our evaluation of *A* (the general case). I call this 'the association-dependence test': if we had no sense that there could be cases like the special case, would there still be *P* in the general case? If so, then it is the general case that is explanatory; if not, then it is the special case.

For the borrowed-lustre cases just described, suppose we had no sense that there could be squares – would square-like objects evoke the same sort of geometric veneration? Suppose we had no sense that there could be live animals as well as stuffed ones – would stuffed animals still evince respect? Suppose we had no sense that there could be factory-farmed meat – would meat in general still merit the same sort of moral avoidance? And so on. In each of these cases, the association-dependence test suggests that it is the particular rather than the general feature that plays the explanatory role in our evaluation of the case. By contrast, applying the test to the boat-sinking case produces no such result: suppose we had no sense that there could be objects weighing exactly 20 pounds – would objects weighing at least 20 pounds still cause the boat to sink?

V.2. *The non-irrationality of associative valuation*

Before turning to the question of how the association-dependence test fares in the case of personal identity, I need to fend off quickly a worry about rationality. If I would not treat non-factory-farmed meat with moral disapprobation in the case where I am unaware of the existence of factory-farmed meat, how can it be rational for me to do so in the actual case?

The answer relies on the assumption that rationality for finite beings allows us to navigate the world on the basis of imperfect rules. Because of

our finitude, it is not irrational to treat one subclass of cases like a larger class of cases, or *vice versa*, even when the subclass or larger class lacks precisely the feature that determines our evaluation of the case to which it is being assimilated. The cases may be intrinsically identical, making differential treatment practically or theoretically prohibitive. And subsumption of individual actions under more significant rules of action may make it possible for us to achieve outcomes we might otherwise find unachievable. No particular instance of eating a piece of chocolate cake is going to make me overweight, but the best way for me to achieve my dietary goals is to adopt a general rule of no cake-eating; no particular instance of giving in to whining is going to spoil my child, but the best way for me to achieve my parental goals is to adopt a general rule of not acceding to whined requests.[14] And this may be so even if eating this particular piece of cake or giving in to this particular instance of whining would actually promote my ultimate goals more effectively. So too might it be rational for me to respond in the same way to all pieces of meat, even though some of the actual instances may not have resulted from factory-farming, and even though in the counterfactual case, where there is no factory farming, I would treat pieces of meat with a non-factory-farming causal history very differently from the way I treat them in the actual case.

V.3. *Application to the personal identity case*

How, though, does all of this apply to the personal identity case? The suggestion is this: suppose we had no sense that there could be identity, as opposed to mere *R*-relatedness – would there still be such thing as prudential concern? I suggest that the answer is 'No'. The concept of prudential concern is tied up with concepts of fairness, responsibility, justice and rationality. Our views about the sorts of rational and moral obligations we have to ourselves and others, considered as beings who exist through time, rest on the assumption that each of us will have at most one continuer, and that this continuer is someone with whom we will be identical.[15] Disruption of this background assumption would result in disruption of the entire framework by which we make sense of this wide range of concepts.[16] And to

[14] I have been helped in my thinking here by Nozick's discussion of principles in *The Nature of Rationality*, esp. ch. 1.

[15] Although I have focused on fission cases, fusion raises corresponding worries: agency presupposes a tight connection between intention and action which fusion disrupts.

[16] For discussion of these issues, see C. Diamond, 'The Importance of Being Human', in Cockburn (ed.), *Human Beings*, pp. 35–62; C. Korsgaard, 'Personal Identity and the Unity of Agency', *Philosophy and Public Affairs*, 18 (1989), pp. 101–32; C. Rovane, 'Branching Self-Consciousness', *Philosophical Review*, 99 (1990), pp. 355–95, and *The Bounds of Agency* (Princeton UP, 1998); Whiting, 'Friends and Future Selves'; Wolf, 'Self-Interest and Interest in Selves'; cf. Velleman's discussion of dignity in 'Love as a Moral Emotion', *Ethics*, 109 (1999), pp. 338–74.

the extent that prudential concern is interconnected with them, it too would be disrupted.

Evidence for this comes from the fact that even in fission worlds it is unique-relatedness and not mere *R*-relatedness that plays a central role in grounding prudential concern from the fission point onwards. Barring repeated instances of fission, Lefty and Righty will be strictly identical with anyone to whom they bear a relation of rational prudential concern; admitting multiple instances of fission, the entire framework of prudential concern begins to break down.[17]

So, I suggest, *R*-relatedness fails the association-dependence test, while identity passes it: if all cases of continuation were cases of identity, prudential concern for *R*-related continuers would be rational; if all cases of continuation were cases of mere-*R*-relatedness, we would lose our grip on the concept of prudential concern.[18]

V.4. *Exceptions, norms and local adaptation*

Nevertheless, as I have maintained throughout, Parfit is right that if Brainy were to undergo fission, the relation of prudential concern he would find himself bearing to Lefty and to Righty would be rational – even if he knew that he was to undergo fission. What Parfit is wrong about is the explanation of this. The reason why it is rational for Brainy to feel prudential concern for Lefty and for Righty is that being *R*-related is very much like being identical. So the interpretation of the exceptional case is parasitic on the interpretation of the normal cases. To the extent that we are able to account for the exceptional, our accounting takes the form of a sort of *local adaptation*; we maintain our background assumptions about continuation in general, and adapt our standard responses to the case at hand.

There is reason to think this phenomenon of local adaptation is quite general. An example is what has happened to the concept of motherhood in the face of recent technological advances. As it has become possible to implant the egg of one woman into the uterus of another, a previously unnoticed distinction has been drawn between *genetic mother* on the one hand and *birth mother* on the other. Since the cases are exceptional, there has been an effort to 'save' the original concept; and since it is the egg-donor whose

[17] One might think that all I have shown is that what matters is *R*-relatedness plus uniqueness-at-a-time, and not identity as such: see R. Martin, 'Fission Rejuvenation', *Philosophical Studies*, 80 (1995), pp. 17–40. I respond to this objection in my *Thought Experiment: on the Powers and Limits of Imaginary Cases* (New York: Garland, 2000), p. 145, fn. 40.

[18] Cf. J. McDowell, 'Reductionism and the First Person', in Dancy (ed.), *Reading Parfit*, pp. 230–50, at p. 234; cf. also M. Johnston, 'Human Beings', *Journal of Philosophy*, 84 (1987), pp. 59–83, 'Reasons and Reductionism', *Philosophical Review*, 101 (1992), pp. 589–618, and 'Human Concerns Without Superlative Selves', in Dancy (ed.), *Reading Parfit*, pp. 49–79, which have greatly influenced my own views on these matters.

genetic information is carried on to the child, it is the birth mother who has been given the status of 'surrogate'.[19] But were the practice to become widespread, the concept of motherhood would break down entirely. We would no longer have the idea of filial concern for one's mother as such, because there would be no unitary concept of 'mother' that lay behind it. We might well have two similar concepts, filial concern for one's birth mother and filial concern for one's genetic mother; and we might well think these concepts were more similar to each other than either would be to the concept of concern for one's child, or for one's spouse, or for one's sibling. But I think that in such circumstances the concept of maternal–filial concern *simpliciter* would have no application.

Similarly, I suggest, in a world where *R*-relatedness without identity was the norm, there would not be a concept of prudential concern of the sort Parfit needs for his argument to succeed. There might be a somewhat similar concept, such as the concept of concern-for-one's-*R*-related-continuer. But there is reason to think that it would not be the same concept as the one we have, the concept that describes the relation we bear to our future selves. In such a world, Brainy's relation to Lefty in both the single transfer and double transfer cases would indeed contain what matters for concern-for-one's-*R*-related-continuer. But the relations would not contain what matters for prudential concern, because there would not *be* prudential concern in the relevant sense.

Now if we say, as I have been arguing, that *identity* is what matters in the explanatory sense, then we can account for the fact that our concepts might well change in the face of such a global disruption. (Even if only the weaker claim is true, that there *might not be* a concept of prudential concern, my argument can still go through.) But the same option is not available if we say that what matters is *R*-relatedness. If it is (mere) *R*-relatedness that explains our valuation of identity, rather than the other way around, then the global replacement of identity by mere *R*-relatedness should make no difference to the value we place on the relation we bear to our continuers. If we follow Parfit in accepting that what matters for prudential concern in the explanatory sense is not identity, we have no way to account for the fact that prudential concern as we know it might not exist under the conditions I have described. If, however, we maintain that what matters for prudential concern is identity, then we are able to account for such a potential disruption. By properly recognizing the way in which contingent features play a role in the organization of our concepts, and the way in which exceptions of this sort depend on a background of normal cases, we are able

[19] For even more extreme cases, see L.M. Silver, *Remaking Eden* (New York: Avon, 1997), esp. pp. 155–229.

to account for the case with appropriate provincialism, and not, as Descartes puts it, to 'fall into the extravagances of the knights of romances'.

VI. CONCLUSION

VI.1. *Summary*

I began by suggesting that there is a danger to philosophical enquiry which ignores what I have been calling the facts of life. That human beings come into existence only through the predictable sequence of events mentioned in the opening paragraph is one of the background truths against which we organize our concepts. At the same time it seems possible that there could be circumstances (fission is one) where a process that is ordinarily identity-preserving might turn out to be entity-creating. That is, it seems possible that there could be a process with the following character: if it happened in one way (what I have been calling 'the 'single transfer case') it would result in the continued existence of some entity over time; but if it happened in another way ('the double transfer case') it would result in the creation of two new entities. (Of course part of what is at issue is whether it is correct to describe the cases as involving 'the same process', but I trust that despite the sloppiness of my language, my meaning is clear.) If the entities in question are self-conscious, as human beings are, then this possibility raises the following puzzle. To the extent that the process itself would be intrinsically the same in both cases, how could the rationality of one's attitude towards one's continuer depend on whether the process ended up being identity-preserving or ended up producing two new human beings? Presumably one's attitude towards one's continuer would rationally be the same in both the single transfer and the double transfer case. And with this much, I have said I agree.

The question that has concerned me in this article has been the question of what lessons can be drawn from this fact. Parfit contends that from it we can conclude that what makes my prudential concern for myself-tomorrow rational is not the fact that myself-tomorrow will (presumably) be identical to myself-today, but only that the former will be connected to the latter by the right sort of causal process that will result in the right sort of relation of psychological continuity and connectedness. I have tried to show that this conclusion can be blocked. I pointed out that Parfit's reasoning rests on what I have called the intrinsicness principle (for *M*), the implicit endorsement of which can be traced to a failure to see that two relations need not coincide. This failure to see that *RPC* can obtain without *M* (that is, failure

to see that *A*'s prudential concern for *B* may be rational, even if the relation that matters for rational prudential concern does not hold between them) can be traced to two things: (a) the fact that an analogous intrinsicness principle is true for *RPC*; (b) the fact that what I called the necessity principle (that *RPC* cannot obtain without *M*) seems undeniable. But the force of the necessity principle can be traced to a fallacious view of how broadly Mill's method of agreement can be informatively applied. While the method of agreement is a valuable tool when we are concerned with causal explanation, it cannot be straightforwardly employed in certain cases of explaining value. In such cases, the necessity principle is false, along with the intrinsicness principle (for *M*).

What this means is that Parfit's argument shows much less than he takes it to show. It shows only that there are conceivable circumstances where it might be rational to bear a relation of prudential concern towards a continuer with whom one was not identical. But it does not show that identity is not what matters in the explanatory sense.

VI.2. *Larger lessons*

Although most of my argument has focused on a single example, I take my discussion to have general implications. In the case I described, we are asked to consider a scenario in which a pair of features that coincide in all actual situations are separated in imagination, and to make a judgement about which of the two features has conceptual primacy. I have argued that the proper interpretation of the case may be precisely the opposite of what it has generally been taken to be. And I think the reason why its implications have been so misunderstood is this: certain patterns of features that coincide only fortuitously may none the less play a central role in the organization of our concepts. To the extent that imaginary scenarios involve disruptions of these patterns, our first-order judgements about them may be distorted or even inverted.[20]

Syracuse University

[20] I owe thanks to Robert Nozick, Hilary Putnam and especially Derek Parfit for extensive discussion of the original version of this paper, which appeared as a chapter of my dissertation. For comments on early drafts, I am also grateful to Richard Boyd, Michael Della Rocca, Terence Irwin, Mohammed Ali Khalidi, Norman Kretzmann, Scott MacDonald, Elijah Millgram, Carol Rovane, Sydney Shoemaker, Susanna Siegel, Jason Stanley, Zoltán Gendler Szabó, Jennifer Whiting, and extremely helpful audiences at Harvard, Cornell and Syracuse Universities. For exceptionally valuable discussion of more recent drafts, special thanks are due to John Hawthorne, Ted Sider and Zoltán Gendler Szabó.