Topoi
https://doi.org/10.1007/s11245-018-9539-3

Rationality and Future Discounting

Arif Ahmed[1]

Abstract
The best justification of time-discounting is roughly that it is rational to care less about your more distant future because there is less of you around to have it. I argue that the standard version of this argument, which treats both psychological continuity and psychological connectedness as reasons to care about your future, can only rationalize an irrational—because exploitable—form of future discounting.

Keywords Decision theory · Future discounting · Personal identity

Positive time-preference is a bias towards the present or near future at the expense of the more distant future. Most people exhibit at least positive time-preference for fixed monetary sums. For instance, you would prefer \$100 now to \$100 in a year's time.[1]

There are three natural justifications for this. The first is that the nominal interest rate is positive. This means that you could simply invest the \$100 now for a return of more than \$100 in a year's time. So you are strictly better off taking the \$100 now.

The second is the declining marginal utility of money. Every extra dollar matters less to you than the last—this is why rich people spend money on luxuries like taxi-rides that many poorer people *could* also afford but choose not to purchase. If you expect that your wealth will increase over time, an extra \$100 is going to be worth less to you in a year's time than it is now.

The third justification involves uncertainty. Many things *other* than an increase in wealth could happen over the next year that would also make \$100 less valuable to you at the end of that period. For instance, you might die; or the monetary authority might impose currency controls. Factoring this uncertainty into your present decision may suppress the value of the future \$100 relative to the present \$100.

Distinguish monetary positive time-preference from *pure* positive time-preference, by which I mean time-preference for *utility* or well-being under conditions of *certainty*. For instance, suppose that you are offered a choice between two goods (sums of money, episodes of consumption etc.) at an early and a late time respectively; and that you are certain that the late good will be just as valuable to you when you get it as the early good when you get *it*. (So you are certain that you will not die over the period, or at least that death will not affect the good's contribution to your welfare.) If you still prefer the early good then you are exhibiting pure time-preference.

None of the three suggested justifications for *monetary* time-preference could justify *pure* time-preference. Well-being unlike money cannot be invested for a future return, so positive interest rates cannot justify it. Every additional unit of well-being is by definition as valuable as the last, so an expected secular increase in well-being cannot justify it. And you are by hypothesis as certain to enjoy the late good if you choose it now as to enjoy the early good if you choose it now. So uncertainty about the future cannot justify pure time-preference either. Can anything justify it?

Many writers have thought not. For instance, Plato thought that discounting the future in this way involves a confusion something like that between what is small and what is far away.[2] Hume and Pigou have also traced

✉ Arif Ahmed
ama24@cam.ac.uk

[1] Gonville and Caius College, Cambridge, Cambridge CB2 1TA, UK

[1] There is a large economics literature and an increasing philosophy literature on this subject: see e.g. Haugen (1995) (which anticipates the themes of this paper, though its line of argument is quite different); O'Donoghue and Rabin (1999); Frederick et al. (2002). As well as containing a brief introduction both to this topic and to a related puzzle about past-future asymmetries (not discussed here), Suhler and Callender (2012) summarizes a wealth of relevant empirical material.

[2] *Protagoras* 356a–e.

Published online: 20 February 2018

time-preference to some (possibly similar) deficiency in our ways of thinking about the future.[3]

My target in this paper is one argument for pure positive time-preference that seems to have a chance of working. This is Parfit's argument based on personal identity, or rather on what he thinks matters about personal identity. In brief, his view is that what gives me special reason for concern about my future is the degree of psychological connection between me now and me then. Since this degree subsides, I am rational to care less about my more distant future.[4]

I argue that this rationale for time-preference faces a severe pragmatic difficulty that does not seem to have been noticed. Given plausible assumptions about the rate at which connections attenuate, anyone who rationalizes her concern for future well-being on Parfitian grounds is exploitable: that is, will voluntarily take a course of action that not only reduces her overall well-being but strictly reduces it at some time without increasing it at any. So Parfit's considerations could only rationalize a form of time-preference that is in this sense inefficient; and this is at least grounds for regarding the latter as irrational.

The plan of the paper is as follows. Section 1 outlines Parfit's position. Section 2 describes and illustrates the argument that exploitation threatens time-preference unless it takes one specific form. Section 3 argues that the Parfitian argument rationalizes a form of time-preference that does not take this form and which therefore is arguably irrational. Section 4 considers some objections.

1 Parfitian Time-Preference

Parfit writes:

> My concern for my future may correspond to the degree of connectedness between me now and myself in the future. Connectedness is one of the two relations that give me reasons to be specially concerned about my own future [, the other being continuity[5]]. It can be rational to care less, when one of the grounds for caring will hold to a lesser degree. Since connectedness is nearly always weaker over longer periods, I can rationally care less about my further future.[6]

By **connectedness** Parfit means a relation that holds with more or less strength between myself at one time and myself at another. The **degree of connectedness** corresponds to the number of direct connections, of the right kind, between those mental states and events (memories, character traits, experiences, thoughts) of mine that obtain or occur at the earlier and those that obtain or occur at the later time.

For instance, if I see a red car on Tuesday morning and—via the normal mechanism—recall seeing it on Tuesday afternoon, then the causal link between my visual experience on Tuesday morning and the episode of conscious recollection on Tuesday afternoon constitutes one such connection 'of the right kind' between myself on Tuesday morning and myself on Tuesday afternoon.

Connection can involve identity as well as causation. If on Tuesday morning I form the belief that Jones drives a red car, and if I preserve this belief until Tuesday afternoon, again via whatever mechanism is normal for me, then the preservation of this state by itself constitutes another connection 'of the right kind' between myself on Tuesday morning and myself on Tuesday afternoon.[7]

I won't—because Parfit doesn't—attempt to give a sharp criterion for exactly which connections are 'of the right kind'. But I will assume, because he assumes, that we have some way to count them.

Now define **strong connectedness** as the holding of very many such connections between a person at one time and a person at another. At one point Parfit suggests that 'very many' requires the holding of at least half of the number of connections that hold across any ordinary day-long period in the life of nearly every actual person.[8] Clearly strong connectedness need not be transitive: although it holds between me on any day and me on the following day, it probably does not hold between me now and me in very many years' time.

Finally, we can say that a person A_1 at time t_1 and person A_n at the later time t_n (written $t_1 < t_n$) are **continuous** if and only if there is a chain of persons-at-times, A_2-at-t_2, A_3-at-t_3… A_{n-1}-at-t_{n-1}, with $t_1 < t_2 < \cdots t_{n-1} < t_n$, such that each is strongly connected to the next, A_1-at-t_1 is strongly connected to A_2-at-t_2 and A_{n-1}-at-t_{n-1} is strongly connected to A_n-at-t_n.

[3] Hume (1978) [1739], III.ii.7; Pigou (1932), I.ii.3.

[4] For a brief account of other normative arguments in this direction see Frederick at al. (2002), p. 359. The authors of this paper there describe Parfit's argument as 'the most compelling argument supporting the logic of positive time-preference'.

[5] For confirmation that this interpolation reflects Parfit's view, note that just before this passage he writes: 'As I have argued, what fundamentally matters are psychological connectedness and continuity' (1984, p. 313). (For an explanation of continuity, read on.)

[6] Parfit (1984), p. 313.

[7] Parfit writes: 'Besides direct memories, there are several other kinds of direct psychological connection. One such connection is that which holds between an intention and the later act with which this intention is carried out. Other such direct connections are those which hold when a belief, or a desire, or any other psychological feature, continues to be had' (1984, pp. 205–206).

[8] Parfit (1984), p. 206. He adds that we might need to weight the connections, for instance by their distinctiveness (ibid. 515 n. 6).

Although strong connectedness is not transitive, continuity, which is just the transitive closure of strong connectedness, *is* transitive; and it holds between myself now and myself whenever I exist at all.

Connectedness and continuity are the two relations that make my future self of special concern to my present self. And although continuity does not come in degrees, connectedness plainly does. Parfit's argument is then that it can be rational to care less about (the well-being of) myself in the distant future than about myself in the nearer future, because time diminishes the degree of connectedness. There are fewer connections between me now and my distant-future self than between me now and my near-future self. That is his normative justification of what I am calling pure time-preference.[9]

2 Rational Time-Preference is Delay-Consistent

This second section outlines a well-known constraint upon time preference that threatens to undermine Parfit's argument. Section 3 realizes that threat: but in this section, I'll just describe the idea and then illustrate it. I'll do this briefly and informally, since the basic point is well-known (for details see Appendix 1).

The representative time-preferring agent (call her Alice) is willing to make trade-offs that a temporally neutral person (call him Bob) would decline. For instance, let T_1 and T_2 be real numbers such that Alice presently cares more about her well-being T_1 units of time (say, days) from now than after T_2 units of time from now. (In the cases that will matter $T_1 < T_2$, but we needn't assume this here.) Then Alice currently prefers waiting T_1 days for one additional unit of well-being to waiting T_2 days for the same benefit. On a natural assumption,[10] there is therefore some $r < 1$ such that she now prefers waiting T_1 days for r units of well-being over waiting T_2 days for 1 unit of well-being, and so would trade in the second for the first; whereas Bob—who is indifferent between early and late well-being—does not have this preference for *any* $r < 1$, and so would never trade in the second for the first.

Focus now on the **delay** between T_1 days in the future and T_2 days in the future, which I'll define simply as the ordered pair (T_1, T_2), where $T_2 > T_1$. The **length** of this delay is $T_2 - T_1$. Its **discount value** $V(T_1, T_2)$—relative to Alice—is the r such that Alice is indifferent between r units of well-being T_1 days from now and one unit of well-being T_2 days from now. (More-or-less equivalently, it is the *smallest* r such that Alice would accept the former in exchange for the latter.)

The key feature that arguably constrains rational time-preference is then:

Delay-consistency (DC): Delays of the same length have the same discount value.

DC constrains the value of a delay to depend only on its duration and not on its futurity. For any T_1 and T_2, $V(T_1, T_2)$ depends *only* on their difference $T_2 - T_1$ and not in any other way on the absolute values of T_1 and T_2. If Alice conforms to **DC** then she must, for instance, value a week's delay in a week's time—that is, (7, 14)—at the same rate as a week's delay in 3 weeks' time, i.e. (21, 28). If Alice violates **DC** then I'll say that she is **delay-inconsistent**.

DC places the following constraint on the trade-offs that Alice is (now) willing to make. Suppose Alice is willing to exchange 1 unit of well-being to be realized 4 weeks from now, for r units of well-being to be realized 3 weeks from now; then she is willing to exchange one unit of well-being to be realized 2 weeks from now, for r units of well-being to be realized 1 week from now. More generally, if: (a) she is willing to exchange N units of well-being to be realized after T_2 for rN units to be realized after T_1; then: (b) she is willing to exchange N units of well-being to be realized after $T_2 + t$ for rN units to be realized after $T_1 + t$, for any t such that $T_1 + t$ and $T_2 + t$ are both positive.

Now suppose that at *any* time Alice takes the *same* attitude towards well-being at the same distance in the future: if e.g. on 1 January she is indifferent between 1 additional unit of well-being on 1 January and x additional units on 16 January, then on 12 January she is indifferent between 1 additional unit on 12 January and x additional units on 27 January. This assumption is formally reflected in the fact that the discount value function V is a function only of a temporal interval measured from the present; it is not also sensitive to which time *is* now present. The usual name for this is **stationarity**.[11]

Stationarity (S): Alice is equally patient at all times.

It is important not to confuse stationarity with delay-consistency. **DC** is a *synchronic* constraint: it says that evaluated at any *fixed* time $t = 0$, delaying a consumption from a future

[9] It does not justify what one might call *dogmatic* time-preference: that is, diminished concern for one's more distant future self *just* because it is more distant in time. Rather it identifies something other than the date whose diminution between now and that date rationalizes a diminution of concern. Parfit writes that the quoted argument 'defends a new kind of discount rate. This is a discount rate, not with respect to time itself, but with respect to the weakening of one of the two relations which are what fundamentally matter [i.e. connectedness and continuity]' (1984, p. 314).

[10] Specifically, we assume that her present value for waiting T_1 days for rN units is continuous in r.

[11] Farmer and Geanokoplos (2009), p. 2.

time $t=t_1$ to the future time t_2 devalues the consumption by a factor that depends *only* on the length of the delay t_2-t_1 and not also on its timing i.e. not in any further way on the absolute value of t_1. **S** is *diachronic*: it says that when evaluated at different times a delay of fixed length and fixed distance from the index of evaluation creates the same discounting of value.

The main point of this section is that if Alice *violates* **DC** but satisfies **S** then she is exploitable. Given a suitable combination of offers she will voluntarily make a sequence of choices that not only reduces her aggregate well-being over time but strictly reduces it at some time without increasing it at any time.[12]

For an unrealistically dramatic illustration, suppose that Alice's delay-inconsistency takes the following form: the discount value of a delay of length 1 day is 0.2 if the delay begins 1 day in the future, but 0.8 if the delay is 2 days in the future. Stationarity implies that she retains this pattern of preference over the next week; so for any N we have:

(1) On Monday she is indifferent between N units (of well-being) on Thursday and $0.8N$ units on Wednesday.
(2) On Tuesday she is indifferent between N units on Thursday and $0.2N$ units on Wednesday.

Now suppose that on Monday her well-being schedule for the future is as follows:

(3) 60 units on Wednesday, 60 units on Thursday.

On Monday we offer her (A) the chance to give up 20 of the Wednesday units for 30 additional units on Thursday. (1) implies that she will accept; so by Monday evening she is facing the schedule:

(4) 40 units on Wednesday, 90 units on Thursday.

On Tuesday we then offer her (B) the chance to give up 40 units on Thursday in exchange for 10 additional units on Wednesday. (2) implies that she accepts this offer too. So by the time she gets to Wednesday she is facing the following schedule:

(5) 50 units on Wednesday, 50 units on Thursday.

Comparing (5) to (3), we see that Alice's own preferences have led her voluntarily to accept trades that make her worse off on some days and better off on none. In this sense those preferences are inefficient. The culprit is delay-inconsistency: given **S** (stationarity), she is open to this form of exploitation if and only if she is delay-inconsistent. To get a feel for this, notice that Bob—whose complete temporal neutrality plainly implies delay-consistency—would accept A but *reject* B.[13]

Does this show that delay-inconsistency is irrational? Well, it gives a reason to call it that; and it certainly is *inefficient*. In any case, delay-inconsistent time-preference is in some sense self-frustrating: Alice is cheating herself out of something that matters to her, namely her own future well-being. The next step is to redirect this concern at Parfit's argument for time-preference.

3 Parfitian Time-Preference is Delay-Inconsistent

Parfit does not go into detail on the relation between (a) the diminution of connectedness between Alice now and her future self and (b) the rate at which she discounts the latter. But there are two obvious and natural ways in which (b) might seem to depend on (a). I'll argue that on both, Parfitian time-preference is delay-inconsistent and so arguably irrational. A third proposal does *not* entail delay-inconsistency; but it cannot support the form of time-preference that Parfit is seeking to justify.

Throughout this section I'll work with a very simple model of diachronic connection. Say that a direct connection *decays* when the trace of Alice's present psychology that preserves it is extinguished. For instance, if Alice sees a red car on Monday morning, then there is a direct connection, sustained by her memory of that episode, that lasts until she forgets it: the extinction of her memory-trace is the decay of that connection. Similarly, the survival of any one of Alice's present character traits constitutes a direct connection between her present self and any future version of her that retains it; when the character trait changes, the connection decays.

The model that I'll use treats the longevity of any particular direct connection as a random variable. Random in what way? Well, a simple-minded idea would be to treat the decay of any one connection as a kind of rare event that is as likely

[12] The examples given here modify that in Mulligan (1996). The basic underlying point, that non-exponential discounting is dynamically inconsistent, is due to Strotz (1955).

[13] For the proof of the general claim see Appendix 1. It is worth noting that foresight makes no difference: it can easily be shown that even if Alice can see what choices are coming, she will *still* accept both choices. For an example of how this sort of exploitation might work in practice, consider the fact that many people pay for annual gym memberships that they then underutilize (Della Vigna and Malmendier 2006, though note that the authors of that paper suggest overconfidence about future performance as a possible explanation. I thank Fabio Paglieri.).

to occur at any one time as at any other. We can express that with the following assumption:

> **Uniformity (U)**: If a connection has survived up to t, the probability that it will decay over the next small interval of time is proportional to the length of that interval.

Slightly more formally, this means that each direct connection decays in the next ΔT units of time with probability $r\Delta T$ for some constant r, called the *intensity* of its decay-process. It follows that the longevity of that connection follows an exponential distribution: if it exists now, then the probability that it will exist—i.e. will not have decayed—after a lapse of T units of time is e^{-rT}. Similar assumptions are often used to model temporal intervals between randomly occurring events like births or deaths in a population, arrivals at a queue or calls coming into a telephone exchange.[14]

It is important to bear in mind that **U** is not saying that *different* connections have the *same* probability of decay over any given time-period. The intensity associated with a memory connection might, for instance, be higher than that associated with a character trait. **U** implies only that the survival of *each* connection is a process with *some* fixed intensity.

It will matter to the argument that different connections *do* decay at different rates. Whether this is true of any actual individual is an empirical matter that I can't settle here. But it is plausible enough: nobody would find it strange that vivid memories or strong hopes and fears last longer than, or that certain character traits or intellectual capacities change less frequently than, memories of everyday perceptual experiences. Setting aside the empirical plausibility of **U**, it would in any case be a serious qualification of Parfit's argument that it only supports efficient time-preference for beings whose psychological connections all decay at *one* identical rate.

With **U** in hand I turn to the argument that Parfitian time-preference violates **DC**; as advertised I derive it from two natural proposals about the relation between the diminution of connectedness and that of future-directed self-concern.

3.1 Connectedness Without Continuity

The simplest proposal is what Parfit's wording most strongly suggests. Let *Conn* (T) be the proportion of Alice's present states that are directly connected to herself T days in the future. Then her present discount value for the delay (T_1, T_2) is simply:

$$(6)\quad V(T_1, T_2) = \frac{Conn(T_2)}{Conn(T_1)}$$

I'll say that in this case Alice **simply discounts** her future well-being.

For instance, suppose that on some appropriate measure, 50% of Alice's current mental states are directly connected with herself after a year (T_1), but only 10% of her current mental states are directly connected with herself after 5 years (T_2).[15] If Alice discounts simply, her present value for the delay (T_1, T_2) is $0.1/0.5 = 0.2$. She now considers one unit of well-being in 5 years' time to be worth 0.2 units of well-being in 1 year, and so would trade anything less than the former for anything more than the latter.

Does simple discounting give rise to delay-inconsistency? The following analogy sets out the intuition behind my argument that it does.

Let S be a composite material body consisting initially of N atoms of polonium-212, which decays very quickly, and N atoms of uranium-235, which decays very slowly. At the outset, the decay of polonium atoms makes a big difference to the proportion of original atoms of S that remain. But as time goes on the slow-decaying uranium atoms will dominate: *their* decay will make an increasing contribution to the rate at which the survival-rate of the original atoms declines. And since uranium decays *slowly*, the proportion of original atoms that remains falls more slowly as time goes on. So this proportion will fall by more over any 1-day stretch in the *near* future than over any 1-day stretch in the *distant* future. The decline in the proportion of original atoms over a delay of fixed length is *not* itself constant but depends on how far in the future the delay is supposed to be.[16]

Back now to Parfit. The decay of a single atom of S corresponds to the breaking at some time t of one direct connection between Alice's current mental state and her mental state at (or after) t. For instance, her losing a memory

[14] Proof: let S_T be the proposition that decay does not occur before T units of time have elapsed, and let $F(T)$ be the probability of S_T. Then $1 - r\Delta T = \Pr(S_{T+\Delta T}|S_T) = F(T+\Delta T)/F(T)$. So $F(T+\Delta T) - F(T) = -r\Delta T F(T)$. Letting $\Delta T \to 0$ we have $dF = -rFdT$; so $F(T) = e^{-rT}$.

[15] The measure might involve some weighting of the states, so that the preservation e.g. of certain memory traces makes less of a contribution than does the preservation of certain character traits. The exact details of the weighting won't matter here.

[16] Formally, let h_1 be the half-life of polonium-212 and let h_2 be the half-life of uranium-235. Then after time t the proportion of original atoms remaining in S is $f(t) =_{\text{def.}} e^{-r_1 t} + e^{-r_2 t}$, where $r_1 = \ln 2/h_1$and $r_2 = \ln 2/h_2$. So if an interval of time begins at time t and has duration x, then the ratio, of the proportion of original atoms at the end of the interval, to the proportion of original atoms at the beginning of the interval, is given by: $V(t,\ t+x) = f(t+x)/f(t)$. This function is delay-inconsistent in the relevant sense if V varies with t as $x > 0$ is held fixed, which is true if $\partial V/\partial t \neq 0$. Tedious manipulation shows that $\partial V/\partial t > 0$ iff $(r_1 - r_2)\, e^{-r_1 x} > (r_1 - r_2)\, e^{-r_2 x}$; this condition holds for any x iff $r_1 \neq r_2$ i.e. iff the isotopes decay at *different* rates.

tomorrow constitutes the breaking of a direct connection between herself now and herself tomorrow (and thereafter). Now suppose that over some stretch of time direct connections with her original state break at random in accordance with **U**. But there are two types of connection: those that break frequently, and those that break infrequently. The foregoing argument shows that the proportion of direct connections that she maintains with her current state falls at a *declining* rate as time passes. So if the discount value she attaches to a delay (T_1, T_2) is the ratio of the weight of her present direct connections with Alice-after-T_1 to the weight of her present direct connections to Alice-after-T_2, this value will *not* remain fixed for delays of fixed length but will increase as the futurity of the delay increases. Alice is therefore delay-inconsistent.

Here is a simple if unrealistically stark illustration. Suppose that Alice's current mental 'bundle' consists of two types of state: memories and character traits, with weights W_1 and W_2 respectively. The preservation of a single character trait or of a single memory between now and some future time t constitutes the existence of a single direct connection between herself now and herself at t; the loss of such an item between now and then constitutes the breaking of one such item. Now suppose that both types of connection decay at random, but memories quickly and character traits slowly. For instance, suppose that over an average 1-day period 50% of memory connections and 20% of character traits are lost.[17]

What the foregoing argument shows is that if she is a simple discounter then Alice's current discount value for a delay of 1 day will depend on when that delay occurs: it increases with the increasing futurity of the delay. That is, Alice is at any time more impatient about delays that occur in the near future than she is about delays that occur in the relatively distant future. Formally, her discount value for a delay (T_1, T_2)—that is, the units of well-being at T_1 that she would exchange for an extra unit at T_2—is given by the following formula:

$$(7)\quad V(T_1,\ T_2) = \frac{W_1\, 0.5^{T_2} + W_2\, 0.8^{T_2}}{W_1\, 0.5^{T_1} + W_2\, 0.8^{T_1}}$$

So if e.g. $W_1 = W_2 = 0.5$ then the discount value of a delay of 1 day is $V\,(1,\ 2) = 0.68$ if that delay occurs 1 day in the future; but if the delay occurs 2 days in the future then it is $V\,(2,\ 3) = 0.72$.[18] So simple discounting makes Alice delay-inconsistent: and if she simply discounts at the same rate today and tomorrow then she is liable to exploitation.[19]

3.2 Continuity as an Additive Constraint

The target passage from Parfit mentions *two* relations with one's future self that might matter to one's present self: psychological connectedness and psychological continuity. The last section assumed that degree of connectedness is the only determinant of the rate of self-interested future discounting. But what if continuity also matters?

Continuity as defined in Sect. 1 is, unlike connectedness, an all-or-nothing relation: any future person is *fully* continuous with Alice now or not at all continuous with Alice now. So continuity (or its absence) should make the *same* contribution to Alice's evaluation of future well-being regardless of when in the future that well-being is supposed to be realized.

There are two very simple and obvious ways to incorporate continuity into the discount function. We might treat it either (i) as making a weighted *contribution* to Alice's present value for future well-being, over and above that of connectedness, or (ii) as a *necessary condition* on Alice's presently attaching any value at all to future well-being. I'll discuss (i) in this section, and I discuss (ii) briefly at n. 22. (I am not claiming that (i) and (ii) are exhaustive; but they do exhaust the obvious and natural possibilities. If they don't help then the Parfitian must find something that does.)

We can formalize (i) as follows. Let the variable C^* be sensitive to whether the person who benefits from some episode of well-being at T is psychologically continuous with Alice now. If so then $C^*\,(T) = 1$. If not then $C^*\,(T) = 0$. For continuity to make a fixed additive contribution to her evaluation of a delay in the realization of well-being is for there to exist some constant λ, $0 \leq \lambda \leq 1$, satisfying:

$$(8)\quad V(T_1, T_2) = \frac{(1-\lambda)\, Conn\,(T_2) + \lambda C^*(T_2)}{(1-\lambda)\, Conn\,(T_1) + \lambda C^*(T_1)}$$

In something closer to English: Alice's present value for a unit of future well-being at a future time is a weighted average of two things: first, the degree to which she is psychologically connected with the recipient of that well-being; second, whether the recipient of that well-being is continuous with her. The discount value of a future delay in the realization of well-being is again the ratio of the value of this weighted sum for the later time to its value for the earlier time. I'll call this **λ-discounting**.

[17] Of course these rates of decay are too fast to be realistic. And in the normal run of things we can expect these states to be replaced by others (as may also be true in the analogy involving radioactivity). These points make no difference to the argument.

[18] Nothing hangs on the simplifying assumption that $W_1 = W_2$. As long as we set both parameters to non-zero values the outcome is the same. (For proof see Appendix 2.)

[19] For an argument that reaches similar conclusions about very long-term discount rates—for instance, in relation to the pricing of environmental degradation—see Weitzman (1998).

What motivates λ-discounting is the philosophical intuition that the degree of connectedness with one's present self cannot be the *only* factor of future self-interest. There might well be future persons who are to some degree psychologically connected with Alice but not psychologically continuous with her; on the Parfitian view, these are *different* persons.[20] But it might matter to *her* whether *she* gets the well-being. That is, if well-being is distributed amongst a range of future people who are all equally strongly connected to her now, she ought to attach additional value to the well-being of the one that is identical to *her*. λ measures that premium. Note that I am not endorsing this line of thought but only making room for it: what makes the room is that λ in (8) may be strictly positive, rather than zero as in *simple* discounting. (And $\lambda = 1$ represents the other extreme of temporal neutrality, at least as regards Alice's descendants-by-continuity.)

In any case it is easy to see that λ-discounting creates delay-inconsistency for the same reasons as—and in wider circumstances than—simple discounting. The mathematical case is straightforward and set out in Appendix 2. Here I'll briefly state the intuition: what established the declining rate of decay of the polonium-uranium composite described at Sect. 3.1 also establishes the declining rate of decay of a composite of uranium and polonium with some third, radioactively inert substance.

More explicitly: the preferences of any λ-discounter over schedules of well-being enjoyed by *her* future self are indistinguishable from those of a simple discounter who gives weight λ to cross-temporal connections that *never* decay.[21] The argument that simple discounting is delay-inconsistent therefore applies here too. As Alice looks further into the future, she sees the decay of more connections. So her concern for her more distantly future self rests more on its continuity, and less on its degree of connectedness, with her present self. But since continuity with her present self perishes more slowly than these other connections (for as long as she lives), the decline of her future self-concern must decelerate. So λ-discounting is delay-inconsistent. Specifically we should expect the λ-discounter to care less about a delay of fixed length, the further in the future that delay is. Again, this leads to exploitability.[22]

3.3 Expectation of Continuity

A third proposal deserves brief comment. This is the idea that Alice presently weights well-being T days from now by her present degree of confidence that her present self is continuous with the person who gets the well-being T days from now. Let Cr be the probability function representing Alice's credence or confidence over a suitable algebra of propositions. Then on this proposal, her discount value for a delay is

$$(9)\quad V\left(T_1,\ T_2\right) = \frac{Cr\left[C^*(T_2)=1\right]}{Cr\left[C^*(T_1)=1\right]}$$

This function plausibly satisfies **DC**. To see why, note first that if $T_1 < T_2$ (and setting aside exotic cases), Alice-now is continuous with Alice-at-T_2 if and only if Alice-now is continuous with Alice-at-T_1 and Alice-at-T_1 is continuous with Alice-at-T_2. It therefore follows from (9) and the definition of conditional probability that:

$$(10)\quad V\left(T_1,\ T_2\right) = Cr\left[C^*\left(T_2\right) = 1 | C^*(T_1 = 1)\right]$$

And the latter is equivalent to Alice's present confidence that continuity holds between T_2 *and* T_1. This quantity plausibly *is* a function of $T_2 - T_1$, or at any rate it is if we assume that (a) when a direct connection decays, it is replaced by another with the same likelihood of decay. In that case (10) is simply the complement of the probability that enough direct connections will be broken over a short enough subinterval of (T_1, T_2) as to violate continuity within that period. If the proportion of direct connections that decay at any given rate is constant, which it is if (a) holds, then this probability depends only on the duration of the interval (T_1, T_2) i.e. only on $T_2 - T_1$, and not on its futurity. So (10) represents an attitude towards the future that is delay-consistent.

The trouble with (10) is that continuity considerations alone cannot justify the form of time-preference that had been our target all along, namely positive time-preference under *certainty*. The aim was to justify Alice's present discounting of future goods relative to present goods when she *knows* that the former will be as valuable to her when she

[20] This follows from his assertion that the future people that will be you are exactly those that are psychologically continuous with your present stage at times when nobody else is (Parfit 1984, p. 263).

[21] This is something of a simplification: nobody lives forever, so it may be that at some t there are future persons with whom Alice's present mental state is connected, but none of those persons are continuous with Alice. But this makes no difference to the point, which could equally well be put like this: the λ-discounter is indistinguishable from someone who puts weight λ on a connection that is especially long-lived in comparison with other types of direct connection. This still makes λ-discounting delay-inconsistent.

[22] λ-discounting treats continuity as a weighted *addition* to Alice's present value for future well-being. We might alternatively think of continuity simply as a *necessary condition* on Alice's attaching any value at all to future well-being; formally speaking, this means that C^* enters *multiplicatively* into Alice's value function, so that $V\left(T_1, T_2\right) = C^*\left(T_2\right) Conn\left(T_2\right) / C^*\left(T_1\right) Conn\left(T_1\right)$. But this reduces to simple discounting for stretches of time over which continuity is maintained (and so on Parfit's hypothesis, over all of Alice's lifetime), and so the objections raised at Sect. 3.1 will apply here too.

gets them as are the latter when she gets *them* (i.e., now).[23] On the Parfitian equation of personal identity with continuity, this means that we are supposed to be justifying the discounting of goods received by a future Alice with which she *knows* that her present self is continuous.[24] So from that perspective, the target is to justify positive time-preference even in the situation where $Cr\,(C^*(T)=1)=1$, for all relevant *T*. *That* form of time-preference is plainly inconsistent with (10); so, it cannot be justified by considerations that justify time-preference only for stretches of the future over which continuity with the present has a positive chance of being broken.[25]

3.4 Comments

That concludes my main argument. Simply put, it is this. Delay-inconsistency is plausibly irrational. Parfit's rationalization of temporal discounting implies delay-inconsistency. Therefore, Parfit's rationalization of delay-discounting fails. The next section examines some objections to it; but first two comments on its scope.

First: the argument raises a problem for Parfit's rationalization on the basis that it makes the rate of future discounting an aggregate of psychological variables that diminish at different rates: some fast, some slow. But nothing in the argument crucially depends on the fact that these variables characterize the *objects* rather than the *subject* of future-directed concern.

We get an argument with that second effect if we suppose that human decision-making aggregates the *output* of two separate decision-making modules, one ('System 1') being impatient and the other ('System 2') patient. This idea, nowadays called the two-systems hypothesis, has roots in Plato[26] and probably further back, and it currently enjoys some empirical support.[27] According to it, we can model a decision maker's valuation of delays (i.e. the overall output) as reflecting a weighted average of two attitudes, namely the 'impatient' and 'patient' discounting functions associated with System 1 and System 2 respectively. If each *individual* system conforms to delay-consistency, the aggregate output—the decision maker herself—will by the foregoing argument be delay-*inconsistent*.

The two-systems hypothesis therefore renders the decision-maker susceptible to exploitation by means of the procedure described in Sect. 2: when faced with an appropriate schedule the decision-maker will make choices that render her worse off by her own lights at all periods than she might have been had she chosen otherwise. (And—if it means anything to say this—*both* of her sub-personal systems will *also* be 'worse off by their own lights' than each might have been had she chosen otherwise.) None of this means that aggregative models should be discarded as descriptive accounts—on the contrary, they enjoy some empirical support and may indeed explain our *actual* future-directed attitudes, which as is well-known do display delay-inconsistency.[28] I mention it merely to illustrate that as well as its present normative application the main argument also has descriptive application.

Second: the argument did not rely on the fact that in Parfit's view the relevant cross-temporal connections are between *mental* bundles. They might connect *any* aggregates some of whose elements decay or are replaced at different rates from others. So again, the argument would apply against anyone—if there is anyone—who tried to justify future discounting in terms of the diminution of some other aggregate of cross-temporal connections. (For instance: if some genetic lines have a greater chance of extinction than others, one could not justify an efficient *social* discount function on the basis that the value to present society of future welfare is the proportion of presently-living people who get, or whose descendants get, to enjoy it.[29])

[23] 'But nobody is ever *certain* about her future survival.' True enough, but pure positive discounting makes a difference even if we allow it. If her confidence that she will survive for at least T units is e.g. 50%, the *pure* positive discounter must value a reward at T at *less* than half of its present value to her. (9) Cannot account for this.

[24] Strictly speaking Parfit does not equate personal identity with continuity but rather with unique continuity (see n. 20). But the equation holds good if as here we ignore exotic cases involving branching and fusion.

[25] Note that on the assumption (a), we can also argue against versions of (i) λ-discounting (Sect. 3.2) and (ii) multiplicative discounting (n. 22) that replace continuity with *expectation* of continuity. More precisely, these theories suppose either (i) $V(T_1, T_2) = (1-\lambda)\,Conn\,(T_2) + \lambda Cr(C^*(T_2)=1) \big/ (1-\lambda)\,Conn\,(T_1) + \lambda Cr(C^*(T_1)=1)$ or (ii) $V(T_1, T_2) = Conn\,(T_2)Cr(C^*(T_2)=1) \big/ Conn\,(T_1)Cr(C^*(T_1)=1)$. Given (a), we have in general that $Cr(C^*(T)=1) = e^{-rT}$, $r>0$. Case (i) and case (ii) therefore both reduce to simple discounting and the argument in Sect. 3.1 apples.

[26] *Phaedrus* 237d, e.

[27] See e.g. McClure et al. (2007).

[28] For evidence that human subjects are delay-inconsistent (i.e. through 'hyperbolic discounting'), see e.g. Kirby and Herrnstein (1995).

[29] It may be worth briefly mentioning one more general point that does follow from the considerations here. Consider reconciliationism: that is, the view (defended in Lewis 1976) that survival is what matters about personal identity, *and* that psychological connectedness is what matters about it. It is plausible that reconciliationism is only viable in a form that implies (a) that one weights consumption of a future stage by one's credence that it and the present stage belong to a single person and (b) that this weight matches the degree of psychological connectedness (as argued in Williams 2014). Clearly (a) and (b) make the position vulnerable to the concerns that this section has outlined.

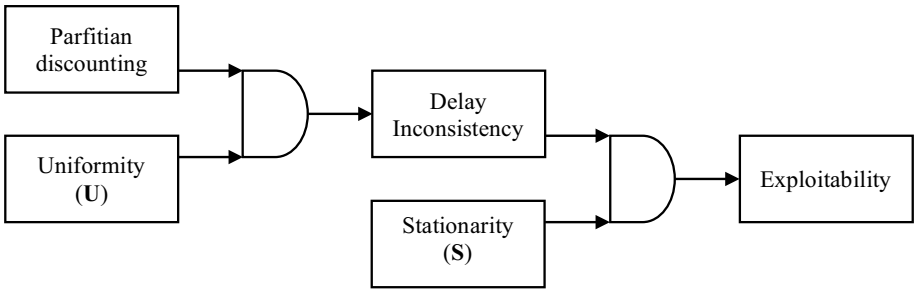

Fig. 1 Logical relations between key claims

4 Objections

The following diagram represents the structure of my argument up to this point. In it, an AND gate represents the conjunction of the claims pointing into it; a claim in a box is entailed by whatever points directly into it (Fig. 1).

As the diagram shows there are two premises that a defender of Parfit might reject. I'll consider these in turn.

4.1 Rejecting S

Stationarity is crucial to the argument because a non-stationary discount function might be delay-inconsistent yet unexploitable.[30] Such a discount function would therefore be immune to criticism from the standpoint of Sect. 3.

Generally, if Alice's time-preference is positive then it suffices to avoid exploitation that the following holds: at any times t, $t+\Delta$ ($\Delta > 0$), and for any delay (T_1, T_2) such that $T_1, T_2 \geq \Delta$, her valuations at t and $t+\Delta$, namely V_t and $V_{t+\Delta}$, satisfy:

$$(11) \quad V_t(T_1, T_2) = V_{t+\Delta}(T_1 - \Delta,\ T_2 - \Delta)$$

And although *stationary* discounting, if it satisfies (11), is delay-consistent and so non-Parfitian for the reasons stated at Sect. 3, Alice could still avoid exploitation if she were non-stationary. If she gets increasingly patient, she may not make the choices amongst future allocations of well-being that caused the trouble in Sect. 3. For instance, suppose that Alice's attitude towards the future evolves as follows:

$$(12) \quad V_t(T_1, T_2) = \frac{1+t+T_1}{1+t+T_2}$$

V_t is delay-inconsistent: it implies that at any time t, the cost of deferring the enjoyment of a future good by a fixed delay $T_2 - T_1$ is a *declining* function of the futurity of that delay. For instance, at $t=0$, the cost of delaying the enjoyment of a future good from tomorrow to the day after is 33% of the present value of having it tomorrow. But at the same time, the cost of delaying a future good from 8 to 9 days in the future is only 10% of the present value of having it in 8 days.

But V_t is also unexploitable, since it satisfies (11). Informally, this is because V_t implies a secular increase in the agent's patience that exactly balances the change in value of future goods as they become less and less future. In consequence, the relative values of a smaller-sooner and a larger-later reward remains constant as the time for the realization of those rewards gets closer: hence the passage of time cannot by itself induce any reversal of preference. The step from delay-inconsistency to exploitability therefore fails in this case.

But non-stationarity is normatively unsatisfactory for reasons that do not apply to Parfit's original idea. For it makes Alice's present evaluation of a *future* delay depend not only on the futurity and length of that delay but also on what date it is *now*. More specifically, her rate of time-preference at any time is a function of the date t—typically a declining function, so that the longer she lives the more patient she gets. But why should her *past* longevity have normative bearing on her *present* concern for her *future* self? Why should it be *a demand of rationality* that (for instance) Alice on 1 January 2016 cares less about her consumption on 10 January than Alice on 1 June 2017 cares about her consumption on 10 June 2017?

Besides, if we *are* in the business of rationalizing non-stationarity, we must give up on Parfit's original claim that the *only* things that are prudentially relevant to present discounting are the continuity and the degree of connectedness between herself now and herself at later times. These quantities do not even appear in (12). It may be that both are the *same* as assessed at 1 January 2017 with respect to

[30] This is a point that the literature on future discounting sometimes misses. For instance, Sozou (1998) argues that preference-reversal occurs if an agent stationed at T discounts any future time t by exactly his subjective probability, at T, of his survival up to $T+t$, where the event of his death is a decay process with unknown intensity λ. Sozou argues correctly that, given an exponential prior for λ, the agent starts out with a hyperbolic discount function. So it is true that the agent's discount function violates delay-consistency. But it does *not* follow, as Sozou claims (p. 2017), that the agent exhibits preference-reversal, because the agent will violate stationarity: as time passes, his estimate of λ will fall, so he becomes increasingly patient.

10 January 2017 and at 1 June 2017 with respect to 10 June 2017.[31] Yet non-stationarity demands a more patient attitude at the later date.

So whilst dropping **S** stops the argument in Sect. 3 from going through, it is doubtful that doing so is either (a) a requirement of rationality or (b) consistent with Parfit's approach.

4.2 Rejecting U

The second assumption was that Alice treats degree of connectedness as the outcome of an aggregate of decay-processes of known intensity, so that the expected number of direct connections that survive between herself now and herself after T is some weighted average of $e^{-r_1 T}$, $e^{-r_2 T}$, … $e^{-r_n T}$, where r_1,…r_n are the known rates at which different kinds of connections decay. This assumption was crucial, for instance, in the illustrative derivation of the delay-inconsistent value function (7) from the assumption (6) that Alice simply discounts the future; more generally it is crucial for deriving the general result in Appendix 2.

Alternative assumptions about the decay of direct connections do *not* have this effect. The simplest and most plausible alternative is that Alice is—like most people—completely in the dark as to the rates of decay of different kinds of connections, but she estimates that overall, the *total* number of connections, between her present self and herself after a lapse of T days, declines at an exponential rate i.e. $Conn\,(T_2)/\,Conn\,(T_1) = f(T_2 - T_1)$ for some exponential function f. If the subject thinks about herself in this way, simple discounting seems to imply delay-consistency, contrary to the inference defended at Sect. 3.1. Moreover, it looks more plausible that subjects *do* think of themselves in this way. In real life nobody even considers, let alone knows, the probability that any given direct connection, or type of direct connection—say, an autobiographical memory—will survive for a day, or a year. In short, the argument at Sect. 3.1—according to this objection—involves an absurd over-intellectualization of what goes on when a real person thinks about his or her future; and when we try to be even slightly more realistic, the whole argument collapses.

There are two replies. The first turns on the distinction—which admittedly I am taking for granted—between the descriptive and the normative. Of course nobody really thinks about these things in the course of actual decision-making. But Parfit's account was supposed to be *normative*—it was saying, not that considerations about direct connections *do in fact* enter into the future-directed and self-interested deliberations of actual people, but that they *ought* to. But if the latter is true then it should hold not only for people who are ignorant of differentials in rates of psychological decay, but also for people who are not. *If* (a) Parfit's normative standard bears on you and me *then* (b) it would also bear on people who differed from us only in respect of knowing the different average rates at which different types of psychological states decay. The objection does nothing to mitigate the doubt that my argument casts on (b); that doubt therefore transfers to (a).[32]

Second, it is *not* necessary for my argument that Alice's knowledge of her cross-temporal psychological connections be as detailed as the objection envisages. If she knows simply that *some* cross-temporal direct connections—e.g. character traits—decay on average more slowly than others—e.g. quotidian autobiographical memories—then the argument already applies. She can see that the degree of connectedness between herself after T days and her present self will decline more slowly as T increases, because as T increases the more stable types of connection tend to make a dominant contribution to the similarity between herself at T and herself now. That by itself is enough to make the Parfitian valuation of future well-being delay-inconsistent as argued at Sect. 3; given the further argument at Sect. 2 it follows in turn that such a valuation is exploitable and so plausibly irrational.

[31] This hypothesis might seem in tension with the argument at Sect. 3, to the effect that the rate at which connections are lost is a declining function of time. But not so. Section 3 exploited the thought that (e.g.) if we look only at direct connections *with Alice's state on 1 January 2017*, the (expected) proportion of the connections among *these* that are lost between 1 January and 10 January 2017 exceeds the proportion of the connections among *these* that are lost between 1 June 2017 and 10 June 2017. It is consistent with this to suppose that the proportion of direct connections with Alice's state on 1 January 2017 that are lost between 1 January 2017 and 10 January 2017 is the same as the proportion of connections *with Alice's state on 1 June 2017* that are lost between 1 June 2017 and 10 June 2017.

[32] A similar point applies to the objection that we ought not to model the survival of a direct connection as a uniform decay process, but rather as governed in some other way that could in principle support the delay-consistency of a Parfitian attitude to the future based on connectedness. For instance, if Alice expects certain severe kinds of psychological disruption to occur at certain specific times in the future then it would be wrong to suppose that e.g. simple Parfitian discounting as in (6) conforms to a rule like (7) over intervals in which those times fall. But if Parfit's justification works at all then it *should* apply to stretches of the future when things *are* expected to go on roughly as normal i.e. in which Alice does not expect any such disruption (or at least, thinks it no more likely to occur at one rather than at another point in this era). Modelling connectedness as an aggregate of uniform decay processes is the natural approach to that situation, because that model distinctively captures the idea that *each* direct connection is as likely to decay in any one small interval of time as in any other of equal length.

5 Conclusion

For any time t in the next 50 years (say), Alice like many other people now has a special concern for just one of the persons who will exist at that future time. It is correct to say that Alice is identical to the object of her special concern at t. But leaving it at that gives us no idea why she should, as most of us do, care *more* about the well-being of that special object's near future than about its more distant future. After all, identity does not come in degrees: she is no *more* identical to herself tomorrow than to herself in 10 years.

The ingenuity of Parfit's idea was that it identifies a relation—degree of connectedness—that comes in degrees, and seems to do so in just the right way to justify caring more about the near than about the distant future. Unfortunately this promising idea cannot rationalize positive time preference: if the measure of connectedness aggregates connections that decay at different rates, then the Parfitian argument supports forms of time-preference that are plausibly irrational because exploitable by means of familiar devices.

Compliance with Ethical Standards

Conflict of interest There is no conflict of interest.

Appendix 1: Delay-Inconsistency and Exploitability

I claimed at Sect. 2 that any delay-inconsistent V is exploitable. To see why, suppose that Alice is delay-inconsistent i.e. for some $\Delta, a > 0, V$ satisfies:

(13) $V(T_1, T_2) = aV(T_1 + \Delta, T_2 + \Delta),\ \Delta > 0, \quad a \neq 1$

This means that she is just willing to exchange one unit of well-being in T_2 days' time for $V(T_1, T_2)$ units in T_1 days' time; and she is just willing to exchange one unit of well-being in $T_2 + \Delta$ days' time for $(1/a)V(T_1, T_2)$ units in $T_1 + \Delta$ days' time. Let the date $t_1 + \Delta$ be (now) $T_1 + \Delta$ days in the future, and let the date $t_2 + \Delta$ be $T_2 + \Delta$ days in the future. And suppose that she is now looking forward to the following schedule of future well-being:

(14) 1 unit at $t_1 + \Delta$, 1 unit at $t_2 + \Delta$

Now suppose first $a < 1$. Then we offer to exchange the unit of well-being at $t_1 + \Delta$ for an additional $a/V(T_1, T_2)$ units at $t_2 + \Delta$. Her schedule now looks like this:

(15) 0 units at $t_1 + \Delta$, $1 + a/V(T_1, T_2)$ units at $t_2 + \Delta$

We now wait for Δ days i.e. until the remaining delay before the date $t_1 + \Delta$ is T_1 days. At that point she is (by **S**) willing to take $V(T_1, T_2)$ units at $t_1 + \Delta$ for each unit at $t_2 + \Delta$. So we choose $\varepsilon, \varepsilon', \varepsilon'' > 0$ s.t. $a + \varepsilon + \varepsilon' = 1$ and $\varepsilon'' = \varepsilon/V(T_1, T_2)$ (these are guaranteed to exist since $a < 1$) and offer her $a + \varepsilon$ units at $t_1 + \Delta$ in exchange for $a + \varepsilon/V(T_1, T_2)$ units at $t_2 + \Delta$. She will accept this deal too, for a final schedule that looks like this

(16) $1 - \varepsilon'$ units at $t_1 + \Delta$, $1 - \varepsilon''$ units at $t_2 + \Delta$

Comparing (16) with (14) we see that Alice has signed up to a sequence of deals that makes her strictly worse off at both $t_1 + \Delta$ and $t_2 + \Delta$ and no better off at any other time by her own lights at any time. That is, at any time she strictly prefers (14) to (16) and yet she voluntarily moves from the former to the latter.

If $a > 1$ we can run a similar procedure in reverse: starting with (14), we offer to exchange the unit at $t_2 + \Delta$ for an additional $V(T_1, T_2)/a$ units at $t_1 + \Delta$. Then after waiting until Δ days have elapsed, we pick $\varepsilon, \varepsilon', \varepsilon'' > 0$ s.t. $1/a + \varepsilon + \varepsilon' = 1$ and $\varepsilon'' = \varepsilon V(T_1, T_2)$ (these are guaranteed to exist since $a > 1$) and offer $\frac{1}{a} + \varepsilon$ units at $t_2 + \Delta$ in exchange for $(1/a + \varepsilon)V(T_1, T_2)$ units at $t_1 + \Delta$. Again, Alice will accept both offers for a final schedule:

(17) $1 - \varepsilon''$ units at $t_1 + \Delta$, $1 - \varepsilon'$ units at $t_2 + \Delta$

This again represents an uncompensated loss that is unavertable even if foreseen.

Appendix 2: Weighted λ-Discounting and Delay-Inconsistency

At Sects. 3.1 and 3.2 I claimed but did not prove that delay-inconsistency attends λ-discounting regardless of how one assigns (non-zero) weights to continuity or to types of psychological connections that decay at different rates, if each connection satisfies **U**.

Here is the proof. Formally, the statement to be proved is:

(18) Suppose that $\lambda_1, \ldots \lambda_n$ are all non-zero and that $\sum_{i=1}^{n} \lambda_i = 1$. Let $V(T_1, T_2) = \sum_{i=1}^{n} \lambda_i e^{-r_i T_2} / \sum_{i=1}^{n} \lambda_i e^{-r_i T_1}$ where $r_i \geq$

0 for all i, j and $r_i \neq r_j$ for some i, j. Then V is delay-inconsistent.

The case that $r_i > 0$ for all i corresponds to weighted simple discounting. The case that $r_i = 0$ for some i corresponds to the requirement that the agent gives some weight to continuity independently of connectedness.[33] (18) covers both cases.

The proof is as follows. Write $f(x) = \sum_{i=1}^{n} \lambda_i\, e^{-r_i x}$ so that for any $x,\ d > 0,\ V(x,\ x+d) = f(x+d)/f(x)$. Then

$$(19)\quad \frac{\partial}{\partial x} V(x,\ x+d) = \frac{f(x)f'(x+d) - f'(x)f(x+d)}{f(x)^2}$$

Expanding the numerator of (19) gives:

$$(20)\quad \sum_{i=1}^{n} \lambda_i r_i e^{-r_i x} \sum_{i=1}^{n} \lambda_i e^{-r_i(x+d)} - \sum_{i=1}^{n} \lambda_i e^{-r_i x} \sum_{i=1}^{n} \lambda_i r_i e^{-r_i(x+d)}$$

And with a little manipulation we can see that (20) is identical to:

$$(21)\quad \sum_{i \neq j} \lambda_i \lambda_j e^{-r_i x - r_j x} \left(r_i e^{-r_j d} - r_j e^{-r_j d} \right)$$

This is easily seen to be the same as (22), which is in turn identical to (23) and hence also (24):

$$(22)\quad \frac{1}{2} \sum_{i \neq j} \lambda_i \lambda_j e^{-r_i x - r_j x} \left(r_i e^{-r_j d} + r_j e^{-r_i d} - r_i e^{-r_i d} - r_j e^{-r_j d} \right)$$

$$(23)\quad \frac{1}{2} \sum_{i \neq j} \lambda_i \lambda_j e^{-r_i x - r_j x} \left(r_i - r_j \right) \left(e^{-r_j d} - e^{-r_i d} \right)$$

But $\left(r_i - r_j \right) \left(e^{-r_j d} - e^{-r_i d} \right)$ is *always* non-negative; it is positive if $r_i \neq r_j$. It follows that every summand in (23) is non-negative and that (23) itself is positive if any two connections decay at different rates. (20) Is therefore also positive in these circumstances; so too, then, is $\partial / \partial x\, V(x,\ x+d)$ (since the denominator on the right of (19) is certainly positive). So if Alice is a λ-discounter of any type then she attaches greater discount value to a delay of fixed length the further it is in the future. So she is delay inconsistent on *any* weighting of the connections whose preservation contributes towards her present concern for her future self.[34]

References

Della Vigna S, Malmendier U (2006) Paying not to go to the gym. Am Econ Rev 96:694–719

Farmer JD, Geanokoplos J (2009) Hyperbolic discounting is rational: valuing the far future with uncertain discount rates. Cowles Foundation: Discussion Paper no. 1719

Frederick S, Loewenstein G, O'Donoghue T (2002) Time discounting and time preference: a critical review. J Econ Lit 40:351–401

Haugen D (1995) Personal identity and concern for the future. Philosophia 24:481–492

Hume D (1978) [1739] Treatise of human nature. Ed. with an analytical index by L. A. Selby-Bigge. Clarendon Press, Oxford

Kirby KN, Herrnstein RJ (1995) Preference reversals due to myopic discounting of delayed rewards. Psychol Sci 6:83–89

Lewis DK (1976) Survival and identity. In: Rorty AO (ed) The identities of persons. University of California Press, Berkeley, pp 17–40

McClure SM, Ericson KM, Laibson DI, Loewenstein G, Cohen JD (2007) Time discounting for primary rewards. J Neurosci 27:5796–5804

Mulligan CB (1996) A logical economist's argument against hyperbolic discounting. Department of Economics, University of Chicago. http://home.uchicago.edu/~cbm4/hyplogic.pdf. Accessed 28 July 2017

O'Donoghue T, Rabin M (1999) Doing it now or later. Am Econ Rev 89:103–124

Parfit D (1984) Reasons and persons. OUP, Oxford

Pigou AC (1932) The economics of welfare. Macmillan, London

Sozou P (1998) On hyperbolic discounting and uncertain hazard rates. Proc R Soc B 265:2015–2020

Strotz RH 1955. Myopia and inconsistency in dynamic utility maximization. Rev Econ Stud 23:165–180

Suhler C, Callender C (2012) Thank goodness that argument is over. Philosophers' Imprint 15:1–16

Weitzman ML (1998) Why the far-distant future should be discounted at its lowest possible rate. J Environ Econ Manage 36:201–208

Williams JRG (2014) Non-classical minds and indeterminate survival. Philos Rev 123:379–428

[33] The weight she gives it is $\sum_{r_i=0} \lambda_i$; and $V(T_1, T_2)$ corresponds to the agent's discount value for delays that occur during the existence of some future person who is psychologically continuous with her (i.e. on Parfit's view, during her own lifetime if there is never more than one such person). Here we assume that there is some stretch of future time over which she is certain of continuity with her present self. To model disruptions of continuity (for instance, sudden death) as events with small but non-zero probability, we must replace the constraint $r_i \geq 0$ with the stronger assumption that $r_i > 0$ for all i. Since the proof goes through on the weaker constraint proposed in (18), it also goes through on this stronger constraint.

[34] I presented versions of this paper at a conference on 'New Trends in Rational Choice Theory' at the Munich Centre for Mathematical Philosophy in November 2016, and at the Cambridge Workshop on Artificial Intelligence, Decision Theory and Severe Uncertainty in March 2017. I wish to thank the organizers of these events, Cédric Paternotte and Adam Bales, and my audiences at those events, including the foregoing but also Richard Bradley, Christian List, Fabio Paglieri, Laurie Paul, Richard Pettigrew and Bernhard Salow. I am also most grateful to two referees for this journal for their many searching comments on an earlier draft of the paper.