A Derivation of Exponential Temporal Discounting from Temporal Neutrality

September 29, 2021

How should we discount utility across time (if at all)? Social scientists and policy-makers model intertemporal decisions using discounted utility theory. Invented by Paul Samuelson and Frank Ramsey in the early twentieth century, discounted utility theory is a way of extending expected utility theory to include time preferences, preferences for when outcomes occur in time. The model gets its name from the fact that people tend to prefer positive outcomes to be delivered sooner and negative outcomes to be delivered later–hence the model discounts the utility of outcomes the later they occur. For the sake of simplicity, Samuelson proposed that we discount all outcomes at a constant rate per unit time. In continuous time this schedule of discounting leads to a discount function that takes an exponential form. Samuelson and Ramsey explicitly disavowed interpreting this form normatively. Samuelson wrote that "The idea that such a [mathematical] investigation could have any influence upon ethical judgments of policy is one which deserves the impatience of modern economists" (1937, 161) and Ramsey insisted that discounting utility over time in any form is "a practice which is ethically indefensible" (1928, p.543). Neither wanted to say that we should discount this way.

Yet within a few decades exponential discounting was widely interpreted as the only rational way to discount in time. A result by Strotz 1955 and subsequent work by Koopmans 1960, Lancaster 1963 and others lent the exponential function a normative interpretation.[1] An exponential function, they show, dominates other functions in the sense that a decision maker using an exponential discount function cannot suffer a preference reversal solely due to their time preferences. Any other function leads to potential preference reversals and hence possible exploitation, a cardinal sin in economics and rational choice theory. The threat of exploitation put exponential discounting on a normative pedestal. As Loewe summarizes, "After Strotz' contribution, the choice of exponential discounting was not an arbitrary choice anymore, nor a choice of convenience; exponential discounting was found to be now the rational standard in intertemporal choice, one based on the fundamental intuition that any normal person is in fact able to plan ahead" (204). Although it has been criticized from various directions, it's fair to say that

[1] See Loewe 2006 for an excellent history.

the exponential discount function is not only widely employed in public policy but also typically understood normatively, as the way one *should* discount in time.[2]

Given its widespread use, it's important to scrutinize the case for a normative understanding of exponential discounting. Put in axiomatic form, exponential discounted utility becomes essentially a representation of the axioms that give us expected utility theory plus a condition known as Stationarity (Fishburn and Rubinstein 1982). Psychologists and behavioral economists who survey people's actual preferences have for a long time reported that they do not tend to satisfy Stationarity. Is Stationarity normatively required? On the basis of Strotz's result, it is widely thought that violations of Stationarity lead to preference reversal and therefore that people tend to discount sub-optimally. This interpretation isn't quite right, however, for Stationarity concerns preferences at only one evaluation point in time. Violating this condition is not a preference *reversal*, as a reversal is a dynamic process that takes time. In recognition of this fact, a violation of Stationarity is sometimes awkwardly dubbed a "static reversal" in opposition to a "dynamic reversal." The best that one can say (as we'll see) is that violating Stationarity "sets one up" for preference reversal, not that it constitutes preference reversal.

Minus the direct connection to preference reversal, Stationarity loses its normative grounding, and with it, so does exponential discounting. If Stationarity had independent normative purchase on us, this might not be a problem. We could still derive exponential discounted utility from it (plus the usual assumptions behind expected utility theory) and retain exponential discounting's normative interpretation. Yet Stationarity doesn't really have much intuitive normative pull on its own. Compare with the axioms of expected utility theory. These have all been contested in one way or another but most have a pretty strong prima facie normative claim on us. Say what you will about the condition that our preferences be transitive, but most will admit that transitivity seems normatively important. Not so with Stationarity.

With that brief set-up, I can now state the aim of this paper: I offer a new argument for the normative status of Stationarity, and by extension, exponential discounted utility. The argument takes as its foundation the philosophical thesis known as *temporal neutrality*. I derive Stationarity from a sharp form of this thesis. Crucial to the argument is distinguishing two forms of temporal neutrality and noticing what is needed to derive Stationarity. One can certainly contest the resulting argument. I'll point to some possible objections in the Discussion. Nonetheless, the argument is valid, novel and based upon independently accepted normative premises; more than that, I feel that it captures the "spirit" behind the imposition of Stationarity. And seeing the argument laid out cleanly allows us to better understand what might make it objectionable.

[2]Famously, when psychologists and behavioral economists survey people's preferences, it turns out that they tend not to discount exponentially. This finding led to the development of scores of descriptively better models using so-called hyperbolic discount functions; see Thaler 1981, Loewenstein and Prelec 1992, Urminsky and Zauberman 2016 for different points of entry into the experimental literature. By itself this empirical finding doesn't directly challenge the normative interpretation of exponential discounting. After all, we often fall short of our normative standards. But there are some direct challenges in the literature, such as Galperti and Strulovici 2014, Farmer, Doyne and Geanakoplos 2009, Drouhin 2009, Frankenhuis, Panchanathan, and Nettle, 2016, in social science, and Ahmed 2019, Pettigrew 2020, and Callender 2021a in philosophy.

1 Exponential Discounted Utility and Rationality

Discounted utility theory considers a decision maker who must choose at some time $t = \tau$ from among various paths of consumption. These consumption paths are streams of temporally-indexed goods. Perhaps one is choosing between apples today and the oranges tomorrow versus oranges today and apples tomorrow. Let the vector $x_t =< x_{t1}, x_{t2}, ...x_{tn} >$ represent the amounts of the n instantaneous goods to be consumed at time t. The decision maker wants to maximize her utility function over these vectors of goods, $u(x_t)$. Because she has time preferences, and in particular, discounts the value of temporally distant outcomes, she modifies her utility function with a discount function, $D_\tau(t - \tau)$. This function represents how she would discount at the decision time, $t = \tau$, and it measures temporal distances from the time of evaluation, i.e., by the delay $t - \tau$. The upshot is that the decision maker aims to maximize

$$u_\tau(x) = \sum_{t=\tau}^{t=\infty} D_\tau(t - \tau)u(x_t) \tag{1}$$

where we assume that $D_\tau(0) = 1$ (that we don't discount the present) and that $0 < D_t \leq 1$ (that the discount function discounts).

The model makes many large assumptions. For instance, it assumes that one can separate time preference from utility and also that one's time preferences are insensitive to the type of outcome considered, i.e., that one discounts apples the same way one discounts oranges. But it doesn't impose any particular form upon the discount function apart from the above constraints. To get from discounted utility to the exponential model, one must choose a particular form, namely, an exponential discount function

$$D_\tau(t - \tau) = \left(\frac{1}{1 + \rho}\right)^{t-\tau} \tag{2}$$

where ρ is the so-called discount rate. The continuous time version of (2) is $e^{-\rho(t-\tau)}$ and that is the reason why this form of discounting is dubbed *exponential*. Exponential discounting is special in that it is constant through time. That is, it takes the same proportion away from utility in each time period. Because an exponential discounter removes the same amount from utility proportionate to the amount of temporal distance elapsed, *when* the evaluation moment happens is irrelevant to such a discounter. Whether the present is today, tomorrow or next year doesn't matter, which is why many presentations of exponential discounting often leave the evaluation time τ out of the formula.

As mentioned, neither Samuelson nor Ramsey endowed exponential discounting with normative significance. What did that was the result by Strotz. Strotz asks, "Under what circumstances will an individual who continuously re-evaluates his planned course of consumption confirm his earlier choices and follow out the consumption plan originally selected?" (171). He proves that the exponential discount function – equation 2 above – is the unique function that will lead to time consistent choices. Of course, one need not discount, and that is consistent with this result because not discounting is constant discounting with $\rho = 0$.

Not long after Strotz's result representation theorems were proven for exponential discounting. These theorems offered the same kind of "carrot" and "stick" approach as their counterparts in expected utility theory. In 1947 John von Neumann and Oskar Morgenstern proved their famous utility theorem, a theorem demonstrating that an agent who satisfies their axioms will maximize expected utility. The carrot is their axioms, which are supposed to be independently normatively compelling. One is drawn to the premises they employ. The stick is the bad outcome that results from non-compliance – in this case, a proof that agents whose preferences violate one or more of these axioms will be susceptible to a Dutch book. A Dutch book is a series of bets that will exploit your beliefs and could eventually lead you to ruin. When axiomatized by Koopmans 1960 and Lancaster 1963, exponential discounting must have seemed to have the same standing as expected utility theory. The carrot was again the arguably independently plausible axioms and the stick was Strotz's result that any other discount function would lead to preference reversals and possible exploitation.

In the well-known Fishburn and Rubinstein 1982 system one essentially derives exponential discounting from five axioms. The first four are commonly employed in obtaining a well-defined utility function. So for present purposes all the action concerns the fifth, Stationarity. Modifying the terminology of Halevy 2015 to suit our purposes, consider outcomes $x, y \in X$, whose values are real numbers, and $t, t' \in T$, the set of times, such that $0 \leq t, t'$, and delays $\Delta_2, \Delta_1 \geq 0$. Then a set of preferences is Stationary if at time $t = \tau$ they satisfy

Stationarity $(x, t + \Delta_1) \sim_\tau (y, t + \Delta_2) \iff (x, t' + \Delta_1) \sim_\tau (y, t' + \Delta_2)$.

When an agent with stationary preferences ranks options, her decision depends only on two differences, the difference between the values of the outcomes (x versus y) and the temporal difference or delay between the two outcomes ($\Delta_2 - \Delta_1$). See Fig.1. *When* the outcomes occur is irrelevant to the exponential discounter.

Although it has been widely known for many decades that Stationarity is not descriptively adequate (Thaler 1981), thanks to Strotz's result the condition has lost little of its normative glow. For instance, textbooks in behavioral economics refer to the axioms of Fishburn and Rubinstein, Stationarity included, as (e.g.) the "axioms of rationality for time discounting" (Dhami 2016, 593).

2 Problems

The justification for a normative understanding of exponential discounting faces at least two immediate and linked problems. One, a decision maker who violates Stationarity does not necessarily exhibit dynamic temporal inconsistency. And two, Stationarity by itself doesn't seem to have strong normative pull. So absent the connection to dynamic inconsistency, Stationarity – and exponential discounting – seems normatively unmoored.

The first problem is simple enough to see. To manifest Stationary preferences a decision maker must have two preferences, but these preferences are elicited at the same time, $t = \tau$. There is no reversal. Reversals require two times. Minus a reversal, there is no automatic path to exploitation.

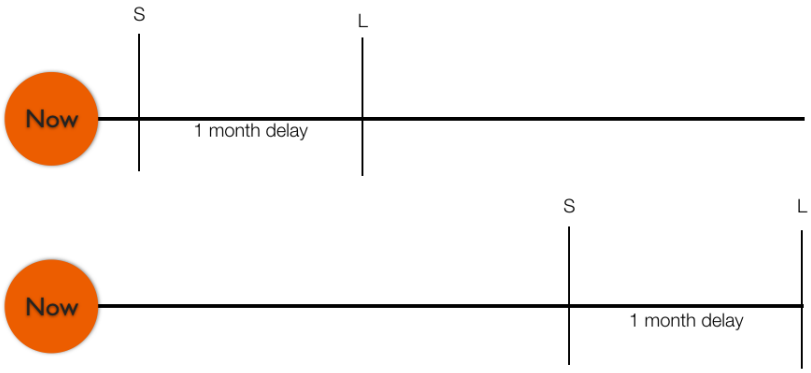

Figure 1: Stationarity: Let the horizontal line represent a tenseless timeline, the dot the evaluation point or Now, S a small reward and L a large reward. A set of preferences that is indifferent between the top and bottom situations is Stationary.

To appreciate the point, compare Stationarity to another temporal condition, Consistency. We can say a set of preferences at times $t = \tau$ and time $t = \tau'$ satisfies Consistency if

Consistency $(x, t + \Delta_1) \sim_\tau (y, t + \Delta_2) \iff (x, t + \Delta_1) \sim_{\tau'} (y, t + \Delta_2)$.

Consistency looks like Stationarity, but note the crucial τ' in the second preference relation. Consistent time preferences mean that one's preferences over temporal outcomes don't change as the present moves from $t = \tau$ to $t = \tau'$, where $\tau' > \tau$. See Fig. 2. Someone who violates Consistency genuinely reverses preferences. In principle that reversal can be exploited. In terms of the figure, the decision maker's preferences change as the "orange dot" – the now – slides along the timeline.

Consider the experimental paradigm typically used in testing our temporal preferences, the smaller-sooner larger-later paradigm. At time $t = \tau$ a subject is asked to decide between a small immediate award of \$100 and a larger award of \$120 a week later. They are also asked to decide between the smaller award and larger award but pushed out a year away and a year and a week away. Studies show that many of us display diminishing impatience (Thaler 1981). We take the small immediate award in the first choice but are willing to wait the week for the larger reward if it is a year away. These non-Stationary preferences are not compatible with exponential discounting. To an exponential discounter, a week is a week and \$20 is \$20, no matter when these occur. But note that having non-Stationary preferences are not enough to be exploited. Suppose someone has the above preference pattern. So long as she sticks to her guns she cannot be exploited. She said she would wait the extra week for the larger reward, and if she still prefers that later, she is not exploitable.

What theorists are implicitly assuming is that she will *not* stick to her guns, that when the smaller reward draws close she will not want to wait the extra week. Of course we do not usually know that because few experiments test the subjects a second time. In

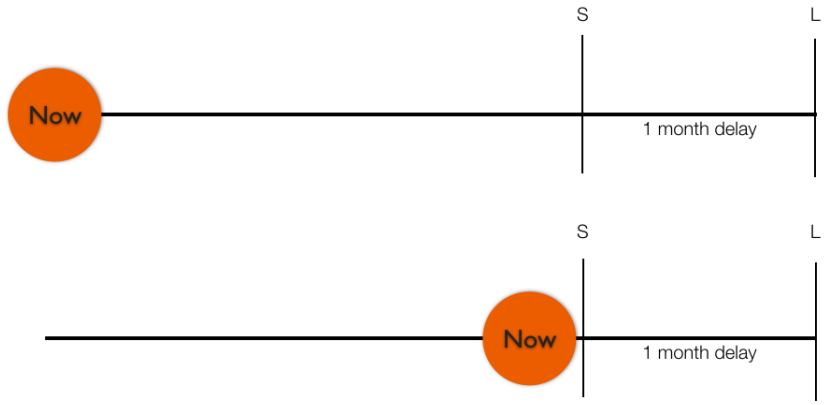

Figure 2: Consistency: Let the horizontal line represent a tenseless timeline, the dot the evaluation point or Now, S a small reward and L a large reward. A set of preferences that is indifferent between the top and bottom situations satisfies Consistency.

fact, in the few recent experiments that have been done that do ask subjects to return to answer more questions, it turns out that many that do switch do not violate Stationarity (about which more below).

The condition that theorists and experimentalists are implicitly assuming is called Invariance by Havely 2015. Invariance acts as a kind of bridge between Stationarity and Consistency. A set of preferences is Invariant if

Invariance $(x, t + \Delta_1) \sim_\tau (y, t + \Delta_2) \iff (x, t' + \Delta_1) \sim_{\tau'} (y, t' + \Delta_2)$

where $t = \tau$ and time $t' = \tau'$. With Invariance, we slide the evaluation point *along with everything else*. It tests whether preferences are indifferent under a time translation that includes the evaluation time. Because the evaluation time moves with the rewards, Invariance tells us whether the decision maker cares about some particular events along the timeline. See Figure 3 and note that the "orange dot" is the same distance from the reward outcomes in both cases.

In Invariance we have isolated what is needed to connect Stationarity to something normatively charged. That is because Stationarity plus Invariance together imply Consistency. (In fact, any two of the conditions imply the third (Havely 2015).) We can prove this using the pictures. Note that each figure (1,2,3) half overlaps with each of the others. Assume that preferences are transitive. Then joining two of the pictures via the common overlap will produce the third picture, for any two pictures. Sliding Stationarity over the overlap of Invariance results in our picture of Consistency. Minus Invariance, however, we lack any path from violating Stationarity to being possibly exploited. And since Invariance is actually not satisfied in a substantial number of subjects when it has been tested (Halevy 2015, Janssens, Kramer and Swart 2017), it is not merely a technical axiom that can be assumed for the sake of convenience. It is a substantial assumption.

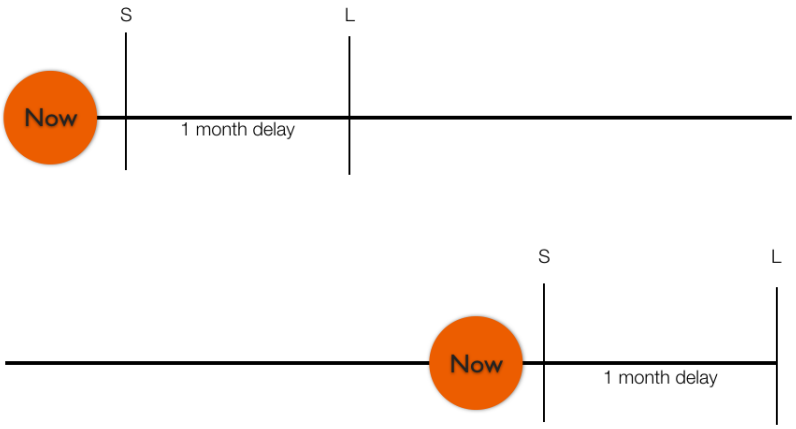

Figure 3: Invariance: Let the horizontal line represent a tenseless timeline, the dot the evaluation point or Now, S a small reward and L a large reward. A set of preferences that is indifferent between the top and bottom situations satisfies Invariance.

This gap between violations of Stationarity and genuine preference reversals leads to the second problem. Minus the connection to Inconsistency, Stationarity on its own just doesn't seem to have much to recommend it normatively. If Stationarity were independently compelling, we could acknowledge the above gap and simply assume that rationality demands it nonetheless. Yet that is not obvious. Stationarity states that if I prefer one temporal stream of outcomes to another, say {eat fish, eat veggies, eat fish} to {eat veggies, eat fish, eat veggies}, then I should also prefer, for any x, {x, eat fish, eat veggies, eat fish} to {x, eat veggies, eat fish, eat veggies}. More aggregate good is supposed to be better. But suppose x=eat fish and I never want to have fish twice in a row? To be fair, this example must not rely on what the first eating of fish will do to the second eating of fish, i.e., make you so full that you can't enjoy the second. The utility must still be instantaneous, so imagine that you get as much enjoyment from the second dish as the first. Still, you may prefer a temporal pattern in your diet, and that doesn't seem irrational. Stationarity assumes that tradeoffs in a time period don't affect overall aggregate goodness. Holding that hardly seems a dictate of reason.

With the link to preference reversals shown to be incomplete and with little independent and transparent rationale, Stationarity becomes normatively unmoored. Can we put it on more secure normative footing? In what follows I want to show that if we understand the philosophical thesis of temporal neutrality the right way, then Stationarity follows as a deductive consequence of temporal neutrality. To see this, we need to distinguish between two kinds of temporal neutrality, both of which are crucial to the argument.

3 Temporal Neutrality and Tense

The philosopher Baruch Spinoza held that

> In so far as the mind conceives a thing under the dictates of reason, it is affected equally, whether the idea be of a thing future, past, or present. (1687, e4:p62).

This philosophical position, known as temporal neutrality, can be traced back to ancient times. The idea is that the rational or prudent person considers well-being across a whole lifespan and not merely at a particular moment. It is advocated by Adam Smith, who implores us to listen to an "impartial spectator" for whom "their present, and what is likely to be their future situation, are very nearly the same" (1790, VI.i.11) and it is developed and defended by Henry Sidgwick 1907. The philosopher John Rawls gives an influential endorsement of this position, holding that

> As Sidgwick maintains, rationality implies an impartial concern for all parts of our life. The mere difference of location in time, of something's being earlier or later, is not in itself a rational ground for having more or less regard for it. Of course, a present or near future advantage may be counted more heavily on account of its greater certainty or probability, and we should take into consideration how our situation and capacity for particular enjoyments will change. But none of these things justifies our preferring a lesser present to a greater future good simply because of its nearer temporal position.(Rawls 1971, 293-4).

David Brink provides a nice succinct statement of temporal neutrality:

> temporal neutrality should be understood to claim that the temporal location of goods and harms within a life has no normative significance except insofar as it contributes to the value of that life. We might say that on this view temporal location has no independent significance or no significance per se" (2015, 358).

Location in time of outcomes isn't by itself a relevant factor when acting rationally. As one can see from the link to rationality, temporal neutrality is an explicitly normative thesis. It is a claim about how best to promote one's well being.

As in the Rawls quote, all involved naturally stress the "independent", "per se" or "by itself" qualifications. Smith, Sidgwick, Rawls, Brink and everyone else who writes on temporal neutrality emphasize that the temporal location *can* rationally matter *indirectly*. It is perfectly rational to take the probabilities of outcomes into account when making a decision. Future events are uncertain. If a future outcome depends on a coin flip landing tails, one should take account of the probability of this happening. Temporal neutrality agrees: one can take the outcome's uncertainty –but not its futurity – into consideration. Because the future is uncertain, it may be that one takes time to be a proxy for uncertainty; but ultimately one is then discounting for uncertainty not time.

Same goes for concerns about mortality, growth in capital, and much more. Another kind of example is taking calendar date to matter. Strotz 1955 gives the example of wanting champagne delivered on one's birthday. If one orders champagne for their birthday, it make sense to want it delivered on the day. A late delivery is valued far less than an on-time delivery. Again, a temporal neutralist can endorse this preference, for what has significance is the birthday, not the temporal location itself.

There is an ambiguity in what temporal neutrality means by 'temporal location' ([blinded reference]). In what type of temporal series is location not supposed to matter? The philosopher John McTaggart 1908 famously distinguished between two temporal series, an *A-series* and a *B-series*. An A-series organizes events via the temporal predicates {past, present, future} whereas a B-series organizes events along a timeline ordered by the earlier or later than relation {earlier than, simultaneous with, later than}. In cognitive linguistics, the distinction is sometimes made between *deictic time* and *sequence time*. Deictic time, like the A-series, has an implicit reference to a deictic center, the now, which is often the time of speaker utterance. Sequence time, like the B-series, is simply calendar or clock time, moments related by a directed ordering relation and typically endowed with a metric that provides a measure of duration. The B-series makes no reference to a now. Both A- and B-discriminations are temporal relations, but A-predicates relate an event to a now, a deictic center, whereas the B-relation refers to two explicitly identified events. Because what is the deictic center changes, statements with A-predicates change their truth value depending upon when they are said, unlike statements with B-predicates. The statement "Nixon was president before Carter" is true at all times, whereas "Carter's presidency is in the future" is now false but was true when Nixon was president. Although it's a slight abuse of terminology, claims invoking A-properties or deictic time are commonly called tensed statements, whereas claims involving B-properties or sequence time are commonly called tenseless statements (see [blinded reference] for references and more discussion).

Disambiguated, we can now distinguish two senses of temporal neutrality:

> *Tensed temporal neutrality*: temporal location in an A-series should have no significance
>
> *Tenseless temporal neutrality* - temporal location in a B-series should have no significance

Tensed temporal neutralism holds that temporal perspective, whether an event is past, present or future, shouldn't matter to you. "When" you are on your timeline shouldn't count in how you value an outcome. If you think of tenses as a kind of temporal indexical, then the idea is that one shouldn't discount for indexical features. That the time is *now* in addition to being (say) noon, GMT, January 1, 2022, shouldn't matter. In philosophy, this type of temporal neutralism is challenged by cases described by Parfit 1985 who shows that we often have a strong desire to discount outcomes when they go past. We sometimes appear to be willing to trade large pain in the past for the elimination of a small amount of future pain, a trade that would decrease well-being along a whole lifetime. Tensed temporal neutralists would deny that this trade makes sense. They

demand a justification for this kind of "past discounting" and so far non-neutralists have had a hard time producing one. For discussion see Brink 2015, Fernandes 2021, Hare 2008 and Suhler and Callender 2012. Regarding the type of discounting considered here, discounting distance in the future, temporal neutralists would be against any type of present or immediacy bias, any kind of intensification of value to an outcome due to its being present or near to the present. Few if any philosophers have defended this type of discounting.

Tenseless temporal neutralists hold that calendar or clock time shouldn't matter. This kind of discounting isn't much discussed in normative theory, and when it is, it is a bit tricky to define. Earlier I gave Strotz's example of wanting champagne delivered on his birthday. Champagne delivered afterwards has less value. Here the position in the B-series matters, the birthday. No one thinks that discounting the value of champagne when it is delivered late is irrational. Tenseless temporal neutralism when unqualified has no advocates. Recall the way temporal neutralism was described by its advocates: it always includes an "independent" or "per se" type clause. To find a defensible tenseless neutralism, then, one must isolate away all of these "impurities" like Strotz's birthday. Suppose we moved his birthday too. Then should he care about that position in time? Arguably not. The problem with this is that it's hard if not impossible to separate pure tenseless preferences from impure ones ([blinded reference], Ziff 1990). To move Strotz's birthday we must also move his actual birth, his friends coming over for the party, and everything else, abstracting away everything but the pure time component. It becomes more or less a re-labelling of the moments of time. The tenseless temporal neutralist holds in effect that this re-labelling should not matter. Greene 2021 agrees that it's hard to distinguish pure from impure preferences, but still claims that so long as a preference is partly grounded in a pure time preference then it's irrational. I'm not entirely in agreement with Greene, but here I just want to establish that there are tenseless temporal neutralists as well as advocates of tensed temporal neutralism.

Advocates of temporal neutrality are not always clear about what kind of time series they mean. Look, for instance, at the Rawls quote. He begins with the B-series talk of earlier and later but then quickly switches to A-series language of present and future. Given the way language works, this imprecision is to be expected. If we had to interpret temporal neutralism as either tensed or tenseless, I think the tensed reading captures what most of them care about most. Sidgwick's concern is to counsel people to not give in to impulsive acts that satisfy the momentary preferences of the present self. It advocates for the importance of now-for-later sacrifices, not 2027 for 2032 sacrifices understood tenselessly. Further evidence for this comes from examining a sophisticated neutralism, such as Brink's, which allows that one may desire one's life to have certain temporal patterns. For instance, maybe you prefer a rags-to-riches life to a riches-to-rags life (Velleman 1991). If so, then you prefer to distribute your resources toward the later moments of your life than the earlier. You're willing to sacrifice earlier-for-later more than later-for-earlier. This pattern is entirely compatible with tensed neutralism because these second-order temporal preferences are all tenseless temporal preferences. At every moment you would endorse this shift of resources toward the later.

In sum, temporal neutralism comes in two forms, each of which has some normative

force. The tensed variety has a long history of distinguished champions; the tenseless variety hasn't been noticed as much, but to the extent it has and can be made clear it also has defenders.

4 The Derivation of EDU

Recall that the conditions Invariance and Consistency together imply Stationarity. Stationarity in turn implies (with the usual axioms of expected utility theory) the exponential form of the discount function. Our derivation is now very simple. It consists of simply noting that tensed temporal neutralism implies Consistency and that tenseless temporal neutralism implies Invariance.

Look again at Consistency. See Figure 2. Consistency says that you are indifferent between outcomes that differ only in where the orange dot is. The orange dot represents your tensed location, the evaluation point $t = \tau$. As "you" change, you still have the same preferences. If you preferred smaller sooner when the now was earlier, you still prefer smaller sooner when the now is later. In other words, Consistency is simply the expression of tensed perspective – location in the A-series {past, present, future} – not mattering to your preferences, which is the very definition of tensed temporal neutralism.

Turn now to Invariance. The shift from the upper preference to lower preference in Figure 3 moves the now *and* the temporal location of the reward outcomes (maintaining the same delay from now). The only thing that isn't changed between the upper and lower conditions is the timeline itself. Invariance states that "preferences are not a function of calendar time" (Halevy 2015, 341). In other words, they are not a function of location in the B-series {earlier than, simultaneous with, later than}, which is the very definition of tenseless temporal neutralism.

To my knowledge, the connections between Consistency and tense and Invariance and tenselessness haven't been noticed before. They seem to be utterly straightforward. Elsewhere ([blinded reference] I show how this observation provides insight into much of what is going on in the exponential versus hyperbolic narrative in behavioral economics. Here I simply wish to point out that temporal neutralism, if understood as the conjunction of its tensed and tenseless forms, implies Stationarity, and therefore, the exponential form of the discount function. *When* outcomes occur is irrelevant to the exponential discounter, as I said, and it turns out someone who doesn't care about A-series position or B-series position will have Stationary preferences.

What is attractive about this derivation of exponential discounting is that it relies on normative principles antecedently accepted by many philosophers. As we saw, Stationarity on its own seemed to have a weak normative basis. Now we can view it as a result of the twin claims that temporal perspective and calendar time shouldn't matter to one's preferences. Of course, the implication is a deduction, so if Stationarity has poor normative standing then so does one or the other of the premises. I'm not suggesting that the implication logically strengthens Stationarity, of course. What I am saying is that Stationarity's normative claim on us before was unclear. Now it is easier to see how it follows from two clear normative principles. This may help us understand why we might

be tempted to endow exponential discounting with normative standing.

5 Discussion

Providing a clear argument for exponential discounting from explicitly normative premises assists us in evaluating the standard model. Here I will not assess the argument in full but note some possible replies and developments for further investigation.

5.1 The First Premise: Consistency

Consistency says that tensed perspective shouldn't matter. This is controversial. As mentioned, Parfit provides examples that lead many philosophers to question this type of temporal neutrality. However, this discussion isn't too relevant to the present debate, for Parfit's examples are about discounting the past. When a painful event like a dentist visit is over, we tend not to care about it as much. Social science rarely deals with this type of discounting (but see Caruso, Gilbert and Wilson 2008, [blinded reference]). If tensed temporal neutralism is false due to Parfit-like arguments, there may be a way of saving this premise by making it about the future and recommending against near-future bumps in value.

So let's focus on challenges to this premise's future-directed form. As I see it, there are basically two kinds of worries one might have.

One might say that possible exploitation doesn't imply irrationality. In the decision theory literature, it is often assumed in the case of belief that sets of beliefs allowing one to be "Dutch booked" imply one is irrational. Whether these arguments show that one is irrational is controversial (Vineberg 2016). Perhaps they only show that one should not enter into bets with bookies who have a creepy amount of information about you. The argument that exponential discounting doesn't and that hyperbolic discounting does allow possible exploitation shares a similar landscape. Suppose we consider an exponential discounter who discounts at a steady 5% at each time step. Compare this person with a hyperbolic discounter who discounts at 5% except for at one time step discounts at 4.99%. The hyperbolic discounter is possibly exploited by once deviating by 0.01%. But what kind of world is that which contains someone ready to pounce on this deviation? If the decision maker has good reason to believe that she doesn't live in such a world, is the hyperbolic form really irrational?[3]

The other type of objection really gets to the heart of the matter: what Pettigrew 2020 calls the *problem of changing selves*. Beginning with Strotz, there has been a large literature on this in rational choice theory, economics and philosophy, so I cannot do justice to it here. Consistency says that as the self – the little "orange dot" in the figures – moves with time, one continues to honor its previous preferences. As you develop through time, your preferences stay the same. If you picked smaller-sooner, then later you still pick smaller-sooner. Yet of course preferences can change. *You* can change. I once preferred

[3]In addition, it's not true that the credal and preferences landscapes really are the same, as Pettigrew 2020 shows. He argues that this difference affects precisely this question; see Pettigrew, section 13.7.4.

chocolate ice cream to coffee flavored ice cream; now I don't, and having previously preferred chocolate isn't a mark in its favor now. Have I done anything irrational in changing? Most would say not, that it's possible to rationally alter your preferences. This change is a natural part of life, and not all of it is irrational. In other cases – consider Ulysses binding himself to the mast because he knows what future-Ulysses under the spell of the Sirens will want – the question gets trickier. There are many responses to the problem of changing selves. Hedden 2015 proposes (but doesn't endorse) retreating to your "ultimate preferences" remaining constant through time. Pettigrew pools preferences in his Aggregate Utility Solution. I cannot survey all of the solutions. We can leave the worry as a challenge: can one plausibly allow for rationally changing preferences and still insist on Consistency? For many methods (e.g., Pettigrew's) the answer will be No. As a result, it may be that we need to restrict exponential discounting's normative status to cases where the decision-maker's preferences are stable.

5.2 The Second Premise: Invariance

The main worry for Consistency is that as *you* change *you* may not share the preferences you once had. The main worry about tenseless temporal neutrality is similar but directed outward. Invariance says that your preferences should remain invariant as *the world* changes and *you* stay the same.

On its face, this condition isn't remotely plausible. Of course calendar time matters. As discussed, there are all those meetings, birthdays, anniversaries and so on to take into account. Zooming out, we're also aging and expecting to die someday. Expecting to be dead in 2080 is an excellent reason to discount the value of personal outcomes then.

I see essentially two ways to respond.

One is to dig in one's heals and try to tease apart pure from impure temporal preferences, as discussed. I find that very unpromising (see [blinded reference], Ziff 1990). A more defensible variant comes from Greene 2021, who argues that so long as a preference is *partly grounded* in a pure time preference it is irrational. He seems to have a vectorial composition of forces picture of preferences in mind. But I suspect the same problem of disentangling pure form impure strikes again at the component level. Compare: I don't want to go to the baseball game downtown tonight. Why? It's too far! Does this mean my preference is partly grounded in a pure bias for the spatially near? Well, the traffic is what's driving the preference. It will take too long and be too frustrating to get there, given the traffic and the distance. I might not care about the distance. If I could be transported Star Trek style to the ballpark, it's a no-brainer for me to go. Is distance a part of the grounds of my preference? I suppose so. If the ballpark were next door and I could walk to the stadium then I'd want to go. But the distance matters only because I could then walk, and walking would be far more pleasant than sitting in traffic. It's not clear that a "pure" component lurks under the preference because the components interact with one another. But perhaps good theory could attribute some aspect of a behavior to a pure component even if its hard to do so in practice? That is what Grüne-Yanoff 2021 suggests, although his concern is to defend Consistency and not Invariance.

Here is a quite different way to respond. Steele 2021 wants to assume that Invariance holds. She argues, and I think she is correct, that this is what Strotz meant to do. Strotz 1955 considers this kind of tenseless discounting, as mentioned, giving the example of wanting champagne on his birthday. The way he models this kind of temporal preference is to time index the utility of outcomes. For this kind of preference, we don't 'model the utility of champagne, full stop; instead we model the utility of birthday-champagne. That has a different value than non-birthday-champagne, for instance. On this picture, we might say time preferences can be usefully modeled as independent of utilities when those time preferences are tensed but not when they are tenseless. Tenseless temporal preferences get absorbed into the utility of an outcome (just as *where* the outcome already is).

I like this response. Like Steele, I think it's hopeless to try to tease apart the pure form impure in our tenseless temporal preferences. More than that, this response puts the focus on the real worry, namely, preferences that flip flop with tensed perspective. This kind of flip-flopping is what worries people (see also Grüne-Yanoff 2021 on this point). Unlike B-time, A-time contains an essentially indexical component to it. Intuitively, why should we change our preferences – rationally – just because the occupant of an indexical changes? Of course, as we saw with the problem of changing selves, there may be reasons. But assuming Invariance arguably puts the focus where it should be. In terms of methodology, we may have to change how we do some experiments, or at least how we interpret them. After all, people often do violate Invariance when it has been tested. Yet there is always the option of focusing only on that population that satisfies Invariance but violates Consistency.

5.3 Getting to Zero

There is a long history in philosophy and early neoclassical economics insisting that one should never discount *at all* purely for reasons of time (Peart 2000, Żuradzki 2016). Translated into the current model, this tradition advocates for $\rho = 0$. Since not discounting is constant discounting and constant discounting is exponential discounting ($e^0 = 1$), the present argument is *compatible with this advice* but *doesn't imply it*. The topic requires more discussion than I can give it here, but it is worth thinking about what additional premises would be needed to get from temporal neutrality to exponential discounting with a value of zero for ρ. The current argument from temporal neutrality constrains the *form* of the discount function but says nothing about the *value* of the discount rate, ρ. For all the current argument cares, one can discount very steeply (large ρ) or not at all ($\rho = 0$); strictly speaking, one could even inflate rather than discount by dropping the arbitrary restriction on negative values ($\rho < 0$).

Representing tensed and tenseless preference types, the conditions of Invariance and Consistency might appear to exhaust the possible forms of temporal neutrality. But that is not so. Notice that both conditions allow that the temporal distance between rewards can legitimately matter. Here I am referring to the delta in the conditions such as $(x, t + \Delta_1) \sim_{\tau'} (y, t + \Delta_2)$. Suppose x is a small reward and y a larger one. In finding the point at which one is indifferent between the two, we are allowing that some values of

the temporal distance Δ compensate for the difference between small and large rewards. That is what leaves non-zero ρ compatible with our two forms of temporal neutrality.

Yet we can imagine a stricter condition, one that insists that our preferences be insensitive to the temporal distance Δ. This can be understood in tensed or tenseless formats; but without Δ the difference between Invariance and Consistency becomes trivial. We end up with a very spare condition, Strict Temporal Neutrality:

Strict-TN $(x,t) \sim_\tau (y,t) \iff (x,t') \sim_{\tau'} (y,t')$

It states that if you're indifferent now ($t = \tau$) between x and y then you should be indifferent at any other evaluation point ($t = \tau'$) too. The temporal relationship between x and y doesn't matter.

To help marshal some intuitions about this condition, let's describe an example in terms of an asymmetric preference as opposed to indifference: suppose that you now prefer buying a red car to a blue car. Strict-TN has two aspects to it, one corresponding to t and the other to τ. First, given the preference for a red car, it demands that you prefer the red car to the blue car no matter when either car is delivered to you. the previous delays we modeled, ($\Delta_2 - \Delta_1$), now don't matter. The blue car might be delivered immediately and the red car in a hundred years. That is of no consequence to the decision-maker satisfying Strict-TN. They want red. The other aspect of Strict-TN is that this preference is consistent over time. It doesn't matter when τ is. In the past, present and future they prefer a red to a blue car. You prefer a life with a red car to a life with a blue car and you maintain this preference throughout your lifespan.

Strict-TN implies that $\rho = 0$. Since the delivery times of outcomes do not matter, there is no room for any non-trivial discounting consistent with this condition. Are there other expressions of TN that imply $\rho = 0$? I'm not sure. Strict TN seems the most natural one that I can produce. It really drives home the idea that time doesn't matter.

Is Strict-TN normatively required? Strict-TN is stronger than the above two conditions, so the worries raised for each of those apply here too. Can one rationally change one's preference from red car to a blue one? And can we maintain a pure versus impure distinction? After all, few would prefer a red to blue car if the red car only arrives after one is too old to drive, or if one is colorblind, or if the red car is a lemon, and so on. Once these problems have been dealt with, one would then have to turn to the question of whether mere "positional" preferences are rationally permissible (see, for instance, Street 2009).

Arguments along this path would have to abandon the moderate Humeanism about preferences that pervades economics and rational choice theory. Hume famously said that "Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger" (1740, 267). Modern economics follows this line of thinking in holding that preferences cannot on their own be criticized. Only sets of preferences can be evaluated for inconsistency. Narrowing discounting down to an exponential form is as far as Humeanism about preferences can go. To get to zero, I suspect, one must abandon the Humeanism.

6 Conclusion

The model of exponential discounting is viewed as the normatively correct way to discount future utility. This has been widely assumed throughout social science and policy for decades. Unlike expected utility theory, it's normative basis has always been a bit shaky. Violating the crucial condition underlying the model, Stationarity, isn't obviously so bad. In the preceding I shine a light on Stationarity that casts it in a normative glow. By observing that temporal neutrality comes in two forms, tensed and a tenseless, and seeing that these forms imply the conditions known as Consistency and Invariance, respectively, I am able to derive Stationarity from independently accepted normative premises. Whether this argument is good enough to be compelling is not clear. But it is a better case for Stationarity's normative role than I've previously encountered and it helps us isolate different kinds of challenges to the standard model.

References

[1] Ahmed, A. 2018. Rationality and Future Discounting. *Topoi*, 1-12.

[2] Brink, D. 2011. Prospects for Temporal Neutrality. In Callender C. *The Oxford Handbook of Philosophy of Time*. Oxford: Oxford University Press, 353-381.

[3] Burness, H. S. 1976. A Note on Consistent Naive Intertemporal Decision Making and an Application to the Case of Uncertain Lifetime. *Review of Economic Studies* 43, 547-549.

[4] Caruso, E., Gilbert, D. and Wilson, T. 2008. A Wrinkle in Time: Asymmetric Valuation of Past and Future Events. *Psychological Science* 19, 796–801.

[5] Dhami, S. 2016. *The Foundations of Behavioral Economic Analysis*. Oxford: Oxford University Press.

[6] Drouhin, N. 2009. Hyperbolic Discounting May Be Time Consistent. *Economics Bulletin* 29, 2552–2558.

[7] Farmer, J. Doyne and Geanakoplos, J., 2009. Hyperbolic Discounting is Rational: Valuing the Far Future with Uncertain Discount Rates. Cowles Foundation Discussion Paper No. 1719. Available at http://dx.doi.org/10.2139/ssrn.1448811.

[8] Fernandes, A. 2020. Does the Temporal Asymmetry of Value Support a Tensed Metaphysics? *Synthese*.

[9] Fishburn, P. and A. Rubinstein, 1982. Time Preference. *International Economic Review* 23, 677-694.

[10] Frankenhuis, W. Panchanathan, K. and D. Nettle, 2016. Cognition in Harsh and Unpredictable Environments, *Current Opinion in Psychology* 7, 76–80.

[11] Galperti, S., and Strulovici, B. 2014. From Anticipations to Present Bias: A Theory of Forward-Looking Preferences, Working Paper, Northwestern University.

[12] Greene, P. 2021. 'Pure' Time Preferences Are Irrelevant to the Debate over Time Bias: A Plea for Zero Time Discounting as the Normative Standard. Response to Callender. *Australian Philosophical Reviews*, forthcoming.

[13] Grüne-Yanoff, T. 2021. In Defense of Intertemporal Consistency. A Discussion of Craig Callender's "The Normative Standard for Future Discounting." *Australian Philosophical Reviews*, forthcoming

[14] Halevy, Y. 2015. Time Consistency: Stationarity and Time Invariance. *Econometrica* 83, 335–352.

[15] Hare, C. 2008. A Puzzle about Other-directed Time-bias", *Australasian Journal of Philosophy* 86, 269–277.

[16] Hedden, B. 2015. *Reasons without Persons: Rationality, Identity, and Time*. Oxford: Oxford University Press.

[17] Janssens, W., Kramer, B., and L. Swart. 2017. Be Patient When Measuring Hyperbolic Discounting: Stationarity, Time Consistency and Time Invariance in a Field Experiment. *Journal of Development Economics* 126, 77-90.

[18] Koopmans, T. C. 1960. Stationary Ordinal Utility and Impatience. *Econometrica* 28, 287- 309.

[19] Lancaster, K. 1963. An Axiomatic Theory of Consumer Time Preference. *International Economic Review* 4, 2210-231.

[20] Loewe, G. 2006. The Development of a Theory of Rational Intertemporal Choice. *Revista de Sociologia* 80, 195–221.

[21] Loewenstein, G. and Prelec, D. 1992. Anomalies in Intertemporal Choice: Evidence and Interpretation, *Quarterly Journal of Economics* 107, 573-597.

[22] Lowry, R., and Peterson, M. 2011. Pure Time Preference, *Pacific Philosophical Quarterly* 92, 490–508

[23] McGuire, J. and Kable, J. 2013. Rational Temporal Predictions Can Underlie Apparent Failures to Delay Gratification. *Psychological Review* 120, 395–410.

[24] Parfit, D. 1984. *Reasons and Persons* (Vol. II). Oxford: Clarendon Press.

[25] Peart, S. 2000. Irrationality and Intertemporal Choice in Early Neoclassical Thought, *Canadian Journal of Economics* 33, 175-189.

[26] Pettigrew, R. 2020. *Choosing for Changing Selves*. Oxford University Press.

[27] Ramsey, F.P. 1928. A Mathematical Theory of Saving. *Economic Journal* 38, 543-549.

[28] Rasmusen, E. 2008. Some Common Confusions about Hyperbolic Discounting. Working Papers 2008-11, Indiana University, Kelley School of Business, Department of Business Economics and Public Policy. Available at SSRN: https://ssrn.com/abstract=1091392.

[29] Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

[30] Samuelson, P. 1937. A Note on Measurement of Utility. *Review of Economic Studies* 4, 155-161.

[31] Sidgwick, H. 1874. *The Methods of Ethics*. London: MacMillan.

[32] Smith, A. 1790. *The Theory of Moral Sentiments*. Oxford: Oxford University Press.

[33] Sozou, P.D., 1998. On Hyperbolic Discounting and Uncertain Hazard Rates. *Proceedings of the Royal Society B: Biological Sciences* 265, 2015–2020.

[34] Spinoza, B. 1687. *The Ethics, Treatise on the Emendation of the Intellect, and Selected Letters*, trans. Samuel Shirley (Indianapolis: Hackett, 1992)

[35] Steele, K. 2021. Why Time Discounting Should be Exponential: A Reply to Callender. *Australian Philosophical Reviews,* forthcoming.

[36] Street, S. 2009. In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters. *Philosophical Issues* 19, 273-298.

[37] Strotz, R. 1956. Myopia and Inconsistency in Dynamic Utility Maximization. *Review of Economic Studies* 23, 165-180.

[38] Suhler, C. and Callender, C. 2012. Thank Goodness That Argument Is Over: Explaining the Temporal Value Asymmetry. *Philosophers' Imprint* 12, 1–16

[39] Sullivan, M. 2018. *Time Biases: A Theory of Rational Planning and Personal Persistence*. Oxford: Oxford University Press.

[40] Thaler, R. 1981. Some Empirical Evidence on Dynamic Inconsistency. *Economics Letters* 8, 201–207.

[41] Urminsky, O. and Zauberman, G. 2016. The Psychology of Intertemporal Preferences. In G. Keren and G. Wu (eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making*, Chichester, West Sussex: John Wiley and Sons.

[42] Velleman, J. D. 1991. Well-being and Time. *Pacific Philosophical Quarterly* 72, 48–77.

[43] Vineberg, S., 2016. Dutch Book Arguments. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2016/entries/dutch-book/>

[44] Ziff, P. 1990. Time Preference. *Dialectica* 44, 43-54.

[45] Żuradzki, T. 2016. Time-biases and Rationality: The Philosophical Perspectives on Empirical Research about Time Preferences. In Stelmach, J., Brożek, B. and & K. Łukasz (eds.), *The Emergence of Normative Orders*. Copernicus Press, 149-187.