The Normative Standard for Future Discounting

Craig Callender

October 1, 2019

1 Introduction

Decisions are typically about outcomes that happen later in time. They demand comparisons of the value of outcomes now versus outcomes later. Should I buy a new car or save for retirement? Have the last piece of cake tonight or tomorrow? Lower carbon emissions now or later suffer greater climate impacts? Intertemporal decisions have triggered hundreds of studies across many fields. Popular subjects include personal finances, addiction, nutrition, health, marketing, and environmental conservation. These studies find that we tend to exhibit positive time preferences; that is, all else being equal, we prefer positive goods, experiences, and states of affairs to be delivered sooner rather than later. Sweets delivered to me tomorrow aren't as desirable as sweets delivered today. Descriptive and normative inquiries tackle *how we make* intertemporal comparisons of utility in such cases and *how we should*. The present paper is about the second issue, the normative question that asks how we ought to translate future utility into present utility. My focus is restricted to individuals and not societies. I want to challenge the conventional wisdom dominating the social sciences and philosophy regarding temporal discounting, the practice of discounting the value of future utility.

Although economists, psychologists and philosophers often sharply disagree on temporal discounting, there is a common picture that many will recognize. Call it the *Standard Model*. It begins with a normative standard. That standard demands that insofar as one is rational one discounts utilities at future times with an exponential discount function. Such a function, as I'll explain, discounts the satisfaction of preferences at a constant rate per unit time. If eating the sweet tomorrow is only worth half as much to me as eating it today, then eating it two days from now should be worth only a fourth as much to me as eating it today, and so on. According to this standard the amount of value discounted should be proportional to the amount of time one waits.[1] Famously, psychologists and behavioral economists inform us that regrettably we in fact discount non-exponentially. The psychologists judge not, but in apparent fit with the normative

[1]Note that this demand is compatible with their being no discounting since $exp[0] = 1$. Some philosophers and neoclassical economists insist on no discounting, as we will see.

standard, they note that non-exponential discounting is often associated with negative outcomes, e.g., borrowing too much, addiction, poverty, and an inability to delay gratification. My sub-optimal time preference is then partly to blame for me eating too many sweets and any health consequences I subsequently incur.

Identified as irrational, a cause or mechanism is sought that explains why we systematically fall short in this regard. Borrowing a familiar "dual systems" picture, it is often assumed that a "hot" cognitive system overrides our "cool" rational systems, bending our exponential functions to non-exponetial in greedy service to the present.[2] We suffer from cognitive biases known as immediacy or present biases. How prone we are to our hot cognitive systems is often assumed to be a personality trait, perhaps even something heritable.[3] That trait may have served our ancestors well when battling saber-toothed tigers, we are told, but it costs us in the modern world as we decide between life insurance and annuities. Cures in the form of interventions are sought that will incline us toward cool rational exponential discounting.

A compelling picture emerges from the chaos of studies across many fields. Apart from a little squabbling, high theory in economics and philosophy inform us about the normative standard. Experiments in psychology and behavioral economics show that we systematically depart from this standard, and that when we do, often these departures are associated with negative life outcomes–just what we might expect from irrational behavior. Moreover, causal mechanisms – e.g., hot/cool systems – linked to the life sciences explain these departures. The Standard Model has it all: theory, data, confirmation, causal mechanisms and explanations all wrapped up in a tidy package.

Despite its attractions, I believe that this model is essentially wrong. The story hinges crucially on the normative standard being correctly identified. Tracing the justification through economics, philosophy and psychology, I'll make what I believe is the best case one can for it, showing how a non-arbitrainess assumption and a dominance argument together imply that discounting ought to be exponential. Ultimately, however, I don't find the case compelling–in fact, I think it's deeply flawed. Non-exponential temporal discounting is often rational–indeed, the paragon of rationality. If this is correct, it's an important point when considering policy interventions. Instead of trying to "fix" non-exponetial discounting because it is irrational and associated with negative life outcomes, we might instead focus attention on why the conditions obtain that make such discounting rational.

Removing the foundation of the conventional wisdom also invites us to reconsider many assumptions in the field, e.g., that time preference is an exogenous parameter of preferences, an intrinsic tendency, or even a personality trait. I will not have space to develop a different picture of what's going on. To a first approximation, however, I suggest we instead understand temporal discounting like we do spatial discounting (see Callender, in preparation). Unfortunately this picture is not tidy. It's downright messy. But if I'm right, it's a more accurate understanding of temporal discounting.

[2]For two different examples, see Metcalfe and Mischel 1999 and Loewenstein and O'Donoghue 2005.
[3]See (e.g.) Anokhin et al. 2014 and Bickel et al. 2014.

2 The Origin of the Normative Standard and the Received View

How the normative standard in the conventional model became the normative standard is an interesting story. Briefly telling it will provide the reader with an introduction to the necessary concepts and arguments.

2.1 Time Preference in Neoclassical Economics

What is the present value of future goods? All creatures who can model the future face this question. But it wasn't until human beings started using complicated financial instruments that the question motivated the development of a science. In the 17th and 18th centuries thinkers such as Johan de Wit, Abraham de Moivre, and Edmund Haley, for instance, worked out the formula for the present value of an annual annuity. It wasn't until the 19th century that philosophers and economists began to characterize *time preference* itself. W.S. Jevons, Irving Fisher, John Rae, Bohm-Bawerk, and Alfred Marshall anticipated much that would later appear in behavioral economics (Loewenstein and Elster 1992).[4]

Like Plato in the *Protagoras* and John Locke in the *Essays* (II, 21), Jeremy Bentham noted that the force of a pleasure or pain varied with its "propinquity or remoteness" (1970, 38-39). People have a tendency to discount the value of future pleasures the more remote they are. Is this tendency rational?

Early neoclassical economists thought not. W.S. Jevon 1871, for example, devised an intuitive formula that incorporated time preference. To maximize total utility across time, he felt, one should distribute goods such that in each time period n the product

$$v_1 p_1 q_1 = v_2 p_2 q_2 = ... = v_n p_n q_n$$

is equal, where v is the marginal utility, p its probability, and q the discount factor given by the fraction of present utility to future utility. Suppose that we have two pieces of cake and are deciding whether to eat them both today, both tomorrow, or one each day. That marginal utility diminishes suggests spreading the cake over the two days. But if uncertainty is high that the cake will still be available tomorrow (low p) and/or if one doesn't much value tomorrow's pleasures (low q), then eating both today might maximize total utility. Note that this model assumes, as contemporary theory does, that one can distinguish v, p, and q.

Like many in this period, Jevons considered p to be a rational factor and q to be irrational. Eliminating q in the above formula represents the allotment

> "...which should be made, and would be made by a being of perfect good sense and foresight. To secure a maximum of benefit in life, all future events, all future pleasures or pains, should act upon us with the same force as if they were present, allowance being made for their uncertainty ... time should have no influence" (72).

[4] In broad outline this section follows Loewe 2006.

Here the concern is not about the form of any function over q – no function is specified – but instead about introducing values of q not equal to unity.

Most in this period agreed that time preference is a kind of character or psychological flaw, one found particularly strongly in the laboring classes (and for Jevons, the Irish (see Peart 2000)). Not valuing the future as the present is "'reckless' and unwarranted, as for instance, when a consumer indulges in a 'drinking bout' instead of buying a new coat" (Marshall 1890, 120). Time preference was seen as the cause of not saving enough, having too many children, and ultimately, poverty. While Fisher mostly agreed with this sentiment, he stands out for admitting that causality could run the other way; namely, poverty tends to exaggerate "the needs of the present" making impatience "partly rational." Which way this causal arrow goes between time preference and personal and social outcomes is still contested. It is a tension intimately connected to the choice of normative standard for time preference.

Gradually the explanations of time preference in terms of character or psychological flaws gave way to what we might regard as more cognitive explanations. Arthur Pigou famously attributed time preference to our mistaken time apprehension – a "faulty telescopic faculty" – making it a kind of cognitive illusion, not something blameworthy; and Frank Ramsey felt that its origin is a "weakness of imagination" (see Frederick et al. 2002). Few felt that temporal discounting is rational.

2.2 Exponential Discounted Utility Theory

Fast forward to the beginnings of modern economics. Expected utility theory has been developed. The setting is now one wherein psychological features are detached from preferences. An outcome x is more valuable to you than x' if you prefer x to x', whether or not x provides more pleasure than x'. Each outcome provides a certain amount of utility to you, and the theory assumes that you wish to maximize your total utility. Like Ramsey 1928, Paul Samuelson wants to add time preferences to this framework, and in 1937 he develops the model that still dominates the theory of intertemporal decision making, *exponential discounted utility theory* (EDU). EDU provides the normative standard accepted throughout most of the social sciences. The basic idea behind it is to modify expected utility theory by introducing a conversion factor that translates future utility into current utility, much as we might apply a conversion factor if translating yen into dollars. On this model, the other country is the future and we convert the currency of the future into that of the present.

The general picture is as follows. At any time, we have preferences for specific outcomes. Each of these outcomes has a certain utility to us; that is, the outcome is valuable to us because it satisfies some of our preferences. Expected utility theory gives us a way of calculating how much value we'll get from a set of outcomes in the presence of uncertainty. The aim is to make a choice that maximizes expected utility.

To isolate the distinctive contribution of time preference, it's helpful to make some assumptions and idealizations. First, ignore uncertainty. We assume that the outcomes contemplated will happen. Second, restrict attention to the *instantaneous utility* of an outcome. If the outcome is the purchase of a car, having that car will affect much else

I later do. We don't want to disentangle all of these contributions, so imagine instead that the outcomes are only valuable at the time they occur – like the taste of a donut immediately eaten. Obviously this is an idealization. Even the memory of tasting the donut can provide value downstream. Third, for simplicity, use only one type of outcome – say, donuts – and fourth, assume that your utility is linear in that outcome. Finally, assume that utility is comparable across different temporal stages of a person't life. Donuts now and donuts later are comparable.

Homer likes donuts. Various states of the world or paths through life will bring him different amounts of donuts at different times, i.e., what economists would call a consumption stream. Let the stream $x =< x_0, x_1, x_2, ... >$ represent an agent's instantaneous utility $u_s(t)$ from an outcome at time t. So x describes Homer having x_0 donuts at time t_0, x_1 donuts at time t_1, and so on. Homer has a positive time preference. He wants donuts, and he wants them now. Viewed from the present, time t_0, he discounts the value of future donuts. For Homer to trade you a present donut for future donuts, you need to give him two donuts tomorrow for one today. Homer is thus applying a discount function $D(t)$ to the utility of tomorrow's donuts. Tomorrow's donuts are worth only half as much to him as today's donuts. Compare: you may need two NZ dollars to exchange for one British pound. The discount function is like an exchange rate between times.

The discount function is a mapping from time to the real numbers. Time can be treated as either discrete or continuous. We don't discount the present moment, so if we let the present be $t = 0$ then $D(0) = 1$. In the case at hand, where Homer values tomorrow's donuts only half as much as today's, $D(1) = 0.5$. The utility Homer derives from the state of the world $x =< x_0, x_1 >$ is therefore expressed as

$$u_0(s) = u_0(x_0) + D_0(1)(x_1)$$

Suppose that a present donut provides Homer 1 utile and he is restricted to two donuts. Then two donuts now maximizes utilty for Homer, giving him 2 utiles; by contrast, one today and one tomorrow yields 1.5 utiles, and none today and two tomorrow yields only 1 utile.

Generalizing, let the current evaluation time be s, then the utility for an agent is given by maximizing:

$$u_s(x) = u(x_s) + \sum_{\tau=1}^{\infty} D_s(\tau)u(x_{s+\tau})$$

where $D_s(\tau)$ is the discount function used at time s for future outcomes at $s + \tau$, where $\tau \geq 1$. This is the *discounted utility* (DU) model.

So far no restrictions have been placed on $D_s(\tau)$. Typically we assume that the discount function *discounts*, i.e., that for $\tau > 0$, $0 < D_s(\tau) < 1$. However, DU is general enough to allow for people whose preferences inflate the value of future outcomes. What turns DU into EDU is a further restriction. The crucial assumption, suggested by Samuelson, is that *discounting is constant through time*. In other words, EDU agents remove the same proportion δ from utility in each time period.

With this assumption in place, one can eliminate the discount function in favor of a discount factor, i.e., let $D(\tau) = \delta^{\tau}$, where τ ranges over the time steps under consideration. Given two timed outcomes, x at τ and x' at τ', then an exponential discounter will prefer x at τ to x' at τ' if $\delta^{\tau}u(x) > \delta^{\tau'}u(x')$. Expressing the total utility through a few time steps, we have:

$$u_{2018}(s) = u(x_{2018}) + \delta u(x_{2019}) + \delta^2 u(x_{2020}) + \delta^3 u(x_{2021}) + ...$$

If Homer values 2019 donuts half as much as 2018 donuts, then he must also value 2020 donuts half as much as 2019 donuts, i.e., one-fourth as much as 2018 donuts, and so forth, if he is to be compatible with EDU.

The discount function and factor are sometimes expressed as

$$D(\tau) = \delta^t = \left(\frac{1}{1+\rho}\right)^{\tau} \tag{1}$$

where ρ is the so-called *discount rate*. Because the continuous counterpart of (1) is $e^{-\rho\tau}$, this is known as exponential discounting. EDU, by itself, does not assume any particular values for the units of time or the discount rate.

The model makes many assumptions, but let's focus on three connected to time. First, as in Jevons, we assume that utility and discounting are independent. EDU makes time preference an exogenous parameter and assumes that we today know the utility of tomorrow's cake to me independent of it being tomorrow. Cakes are cakes on this model, not today-cakes and tomorrow-cakes. By contrast, EDU doesn't pull out a *flavor* function or a space function. We don't say cakes are cakes when it comes to chocolate versus vanilla cakes or close cakes versus distant cakes. Second, the model presumes that discounting is independent of the type of outcome obtaining. D is a function of time, not of x. If Homer discounts the value of donuts by half per year, he also discounts the value of any other outcome by half per year. Third, discounting is assumed to be constant. This assumption is also unique to time. If we arranged goods by weight, we can't find a good per pound constant conversion factor that applies to all goods. In EDU, time possesses a Newton-like equable flow, sweeping over all states of affairs equally, washing over them at a constant rate and leaving their instantaneous utilities unaffected.

Why should we discount according to EDU? Samuelson is writing in the "preferences are preferences" era. That is, preferences are taken as basic input, detached from psychological theorising and left unjudged (only sets of preferences can together be judged for consistency). We are given no explanation of time preference. Nor is any argument provided for treating EDU normatively. Because EDU adopts a simple decay function, it is very easy to work with. Apart from convenience, however, Samuelson says nothing on its behalf. Just the opposite. He highlights the "arbitrariness" (156) of many assumptions and writes that it is "extremely doubtful whether we can learn much from considering such an economic man" (160). Finally, if there was any doubt on the issue he concludes

> any connection between utility as discussed here and any welfare concept is disavowed. 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.

One can't get much clearer than that. So how did EDU obtain its normative force?

2.3 Strotz, Normativity and Axiomatic Foundations

Historically the link to normativity is provided by a "dominance" result by Robert Strotz. Strotz 1956 links EDU to a kind of time consistency. An optimal schedule of consumption is time consistent if it is still optimal when reconsidered at some later time. Put another way, if one's preferences are time consistent, then one can move along a decision tree, maximizing utility at each time step, never having to deviate from an *ex ante* plan. Strotz's article contains a theorem that is widely reported as proving that a schedule is time consistent iff one discounts exponentially. Since one can in principle exploit inconsistency, EDU is thus regarded as dominating all other discounting strategies.

To quickly see the idea, imagine a choice between receiving a \$100 now or \$110 in a week. You prefer the smaller-sooner reward, \$100. Your time preference is such that you're willing to give up \$10 to get the reward now. Now what if you are asked about the same choice but delayed another week?

Supose that you discount according to EDU. Then you apply a constant discount factor to every time period. Say you discount value at 50% per week. When asked about next week, you are comparing \$100 versus $0.5 \times \$110 = \$55$. Since $\$100 > \$55$, you choose the smaller-sooner reward. What about the week after that, still viewed from the same time? Then you are comparing $0.5 \times \$100 = \$50$ against $0.5 \times 0.5 \times \$110 = \$27.50$. Again you'll pick the smaller-sooner reward. For you, smaller-sooner wins each time, and this is true for an exponential discounter no matter how far out we delay the choice.

Suppose instead that your discount factor changes with time. Rewards for next week are discounted by half, just as in the above example, but after that you don't care to discount more. You're the sort of person who doesn't want to wait a week, but having waited one, doesn't much care whether it's one week or two weeks. Then because $\$100 > \$55$, you'll take small-sooner, just as above. But looking out two weeks, you're comparing $0.5 \times \$100 = \$50$ to $0.5 \times \$110 = \$55$. For that choice you prefer larger-later to smaller-sooner. Your time preference causes you to reverse your preference. (Whether this counts as a genuine preference reversal, however, is something we'll discuss later.)

In rational choice theory, preference reversals are judged to be a cardinal sin. The worry is that this flip-flopping can be exploited by others. One can't criticise high discount rates. That's your business. What one can criticise are sets of inconsistent preferences. Your preference for smaller-sooner reverses into a preference for larger-later when nothing has changed but the lapse of time. That is thought to be exploitable and therefore open to criticism, for even by your own lights you should not want to be exploited.

That, in short, is the route to normativity for EDU. Suppose you have a preference between two rewards delivered at two times. Strotz proves that only exponential discounting preserves that preference when the two rewards are delayed by the same amount of

time. All non-constant discounting will eventually produce a time at which a preference reversal will occur, thereby leaving one open to exploitation. As Loewe 2006 nicely sums up, "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).

Before moving on, it's worth pointing out that Strotz's result seems to put EDU on very familiar ground. As I mentioned, expected utility theory was already developed. EDU could be viewed as an extension of that theory when time preference is added. Moreover, expected utility theory was commonly viewed as normatively compelling. In 1947 John von Neumann and Oskar Morgenstern proved their famous utility theorem. This theorem proves that an agent who satisfies various axioms will maximize expected utility. 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 so that you could eventually be led to ruin. The threat of exploitation, plus the idea that each of the axioms had at least a prima facie case in its favor, provides some reason to treat expected utility theory normatively. Thanks to Strotz's result, the situation for EDU looks similar.

In fact, we can put our finger on just how similar. As with expected utility, EDU preference representation theorems were proven, seeking to demonstrate that if some plausible axioms hold then one can be represented as an expected utility maximizer with an exponential discounting function. Scores of such theorems exist, varying in many details (e.g., deterministic versus indeterministic streams, finite versus infinite time). Some of the most well-known include Koopmans (1960), Lancaster (1963), and Fishburn and Rubinstein (1982).

Despite all the differences, it turns out that the crucial axiom underlying EDU in these theorems turns out to be *stationarity*. Fishburn and Rubinstein's system, for instance, employs five axioms. The first four are axioms commonly used to obtain a well-defined utility function. One can certainly object to these, but then one is objecting to something much more general than EDU. The fifth and final axiom, stationarity, is what transforms their utility function into an EDU utility function. Because it will soon play a large role, let's carefully define this notion (using Halevy 2015's definitions). Consider outcomes $x, y \in X$, whose values are real numbers, and $t, t' \in T$, the set of dates, such that $0 \leq t, t'$, and delays $\Delta_2, \Delta_1 \geq 0$. Then a set of preferences is stationary if

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

Ranking two outcomes, the stationary agent has preferences such that the rank depends only on the values of the outcomes (x versus y) and the delay between the two outcomes ($\Delta_2 - \Delta_1$). See Fig.1. Fishburn and Rubinstein prove that there exists a utility function such that

$$(x, t) \succcurlyeq (y, t') \Leftrightarrow \delta^t u(x) \geq \delta^{t'} u(y)$$

given any $\delta \in (0, 1)$ and that an exponential discounter represents someone whose preferences satisfy these axioms. A non-exponential discounter, by contrast, as in the example

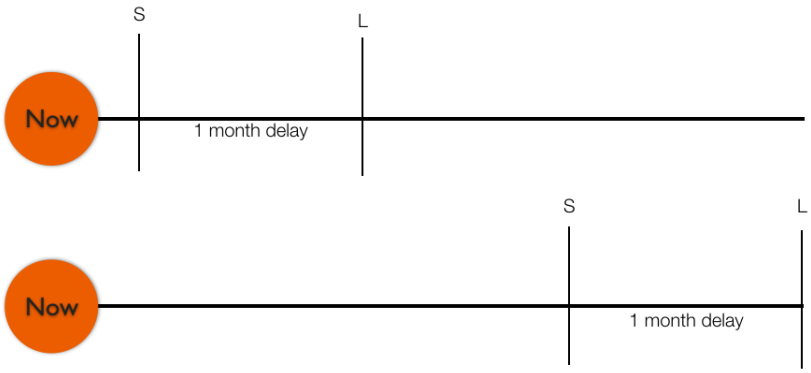

Figure 1: Stationarity: Let the horizontal line represent time, S represent a small reward and L a large reward. An example of stationarity is a set of preferences that is indifferent between the top and bottom situations.

above, is someone whose preferences are not such that they depend only on the values of the outcomes and the delay between them. They might prefer smaller-sooner for next week and larger-later for the week after that, yet in both cases the outcomes and delays are exactly the same.

EDU is thus put on familiar foundations. Normative force originates in the argument that if your preferences violate one of the axioms of EDU, then one is susceptible to exploitation. Self-sabotaging preferences are supposed to be manifestly irrational. Adding extra motivation, the axioms themselves are supposed to each be prima facie compelling, just as in Von Neumann-Morgenstern. Textbooks now refer to the axioms of Fishburn and Rubinstein as (e.g.) the "axioms of rationality for time discounting" (Dhami 2016, 593).

2.4 The Received View

The comparison to expected utility theory becomes complete when we turn to psychology and behavioral economics. Over the past few decades those fields have discovered many well-known examples of *systematic* violations of the probability calculus. For instance, the Allais paradox is a choice of two gambles that seems to violate an axiom of expected utility theory. Subjects are offered a choice between two gambles, both precisely the same from the perspective of expected utility, yet subjects overwhelminglty prefer the gamble that minimizes the chance of receiving nothing. Experimental patterns like these have lead many to think we systematically depart from normative rational behavior.

Same here. Many so-called "anomalies of temporal choice" have been discovered (Loewenstein and Prelec 1992). Perhaps the most pervasive allegedly non-normative pattern is diminishing impatience, illustrated in the last section's toy example. Asked at some time whether they prefer a small immediate reward or a larger later reward, subjects may respond that they prefer the small immediate reward; but asked about

waiting that same interval where that interval is much later, often subjects will prefer to wait and obtain the larger reward. For example we find:

$$(\$100, now) \succ (\$120, 1week)$$

$$(\$100, 1year) \prec (\$120, 1year + 1week)$$

is a common pattern (Thaler 1981). The extra reward is not worth waiting for if soon, but if later it is.

This familiar pattern of preferences, which we might write as

$$u(x) > \delta^{t+\tau}(y)$$

$$\delta^{t'}(x) < \delta^{t'+\tau}(y)$$

is impossible in EDU. In terms of the Fishburn and Rubinstein, these preferences violate a combination of axioms, in particular, Stationarity. In EDU, a week is a week, whether present or a year away, whereas that is not the case to most of us.

Diminishing impatience turns out to be just one of many such anomalies. To accommodate these patterns, a variety of non-EDU models of DU have been developed. Motivated by handling diminishing patience, many so-called *hyperbolic* discount functions are proposed.[5] It's currently an open question whether any hyperbolic function is descriptively adequate to all the known temporal anomalies. Since there are so many such anomalies (see Urminsky and Zauberman 2016 for an excellent review), it's hard to imagine that any simple function will predict all the patterns so far discovered.

To recap, in many empirical studies, reversals of the kind discussed are found to happen again and again. Hyperbolic discount functions predict the existence of reversals, so to some extent hyperbolic functions have been empirically confirmed. Assuming EDU is the normative standard, we should discount exponentially, not hyperbolically. We thus systematically depart from normative behavior. What makes our behavior non-normative is that we can be exploited by these reversals. This sin replaces the concerns of the neoclassical economists, namely, steep discounting (e.g., high ρ). Jevons probably wouldn't be impressed by an exponential discounter. He or she can adopt arbitraily high discount rates and still remain normative. But he or she will not adopt self-sabotaging preferences. That is enough according to EDU.

Further confirming the normative interpretation of EDU are associations between non-EDU discounting and various negative personal and social outcomes. I can't survey

[5] In some contexts all non-exponential forms are referred to as hyperbolic discounting, even if the function isn't strictly hyperbolic. In other settings the denomination *hyperbolic* is reserved for genuinely hyperbolic functions. The following is a common one:

$$U(x) = \sum_{i=1}^{\infty} \delta^i \tfrac{1}{1+\rho i} u(x_i) \tag{2}$$

where in this case the continuous time counterpart of (4) is $1/1+\rho\tau$ and τ is the duration between the consumption and the evaluation point.

what's known here, but there are studies investigating associations between time preference and poor savings behavior, tobacco use, alcohol use, drug abuse, addiction, obesity, risky sexual behavior, and much more. For an entry into the literature and many references, see Story et al 2014.

Just as in the case of empirical violations of expected utility, mechanisms are posited that would explain why we depart from this normative standard. Although they differ in important details, there are many dualling cognitive systems approaches to discounting. In one manifestation of the idea, we have two competing cognitive systems, System 1 and System 2 (Evans 2008). System 1 is evolutionary older. It is a fast system requiring little attention and focused on the present. System 2, by contrast, is a newer system, one requiring more attention and focusing on the future. System 2 might "want" to discount exponentially, but System 1 with its greedy immediate needs bends our discounting curves away from the exponential. In Metcalf and Mischel 1999's version, the contrast is between a "cool" knowledge system and a "hot" emotionally charged motivational system ("know" versus "go"). Shefrin and Thaler 1992 posit a competition between a planner and a doer, where in this case the big difference is between short and long term planning.

We now have the tidy package advertised at the introduction. High theory in economics, rational choice theory and philosophy suggest that flip-flopping is a problem. EDU is picked out as normatively special because it prevents time preferences from causing flip flops. Alas, psychology and behavioral economics convincing shows that we do systematically flop flop. This non-EDU behavior is associated with many negative outcomes. Our time preferences are part of the reason why we don't stick to diets, remain addicted, fail to save enough for retirement, and more. Ultimately these behaviors are blamed on a Manichean battle for our souls generated by evolution.

3 A New Justification for EDU

Strotz's theorem is often glossed as establishing a biconditional between exponential discounting and time consistent preferences. Yet there is a gap between the two, causing a large problem for the standard justification. In this section I'll describe this problem and then propose a fix, describing a new master argument for the normative force of EDU. This argument is novel, valid, and contains clearly normative premises that are independently accepted elsewhere in philosophy and economics. I don't believe the argument is compelling and in later sections I will explain the reasons why. However, I do believe that it is the best argument yet produced for a normative understanding of EDU.

3.1 Problem

Preference reversal is supposed to be the major sin associated with non-EDU preferences. Strotz's result shows that non-EDU preferences are dominated by EDU-consistent preference sets. The key axiom underlying EDU, we saw, is Stationarity. This is also the axiom tested empirically and found to be violated. The case for EDU hangs on non-Stationary preferences being linked to preference reversals.

The problem with this is easy to see: violating Stationarity doesn't reverse any preference. You prefer \$100 now to \$120 in a week but \$120 in a year and week to \$100 in a year. So what? You haven't changed your mind. No preference has been reversed. You simply have a preference for two different outcomes. The preferences between the two outcomes are expressed at one time. Reversals happen across time. No time, no reversal. The literature sometimes dubs violations of Stationarity "static preference reversals," but that just means that they are not genuine reversals.

Maybe Stationarity is normatively required? The problem with this, as many have noted, is that on its own it has very weak normative force, if any. Violating Stationarity can seem eminently plausible. Stationarity states that if I prefer one stream 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 seems 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. 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 one time period don't affect overall aggregate goodness. Holding that hardly seems a dictate of reason. Stationarity on its own has little normative claim on us.

Absent a reason to believe Stationarity is normatively warranted on its own, is there reason to nonetheless accept it? Although there is no preference reversal involved in non-Stationary preferences, there might be something a little awkward about non-Stationary preferences. Can this awkwardness be boosted into an outright problem?

Begin with the vice we want to avoid, preference reversals. Preference reversals indicate a kind of time inconsistency. Let's be clear what this means. Using the same terminology and constraints as before with Stationarity, a set of preferences is time consistent if

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

Consistency looks like Stationarity, but note the crucial t' in the second preference relation. Consistent time preferences mean that one's preferences over temporal outcomes don't change as one moves from t to t'. See Fig. 2. If in 2018 one prefers a large later reward to a small one, then if time consistent one still does when those later times arrive. Violating Consistency is to genuinely reverse preferences. In principle this reversal can be exploited.

An urgent question therefore beckons: can we get from violations of Stationarity to violations of Consistency? If not then the normative standing of Stationarity (and EDU as a consequence) hangs on almost nothing.

It turns out that the answer is quite simple and goes through a condition called Invariance. Halevy 2015 states a beautifully simple relationship amongst the three temporal conditions, Consistency, Stationarity, and Invariance, namely:

> Any two implies the third.

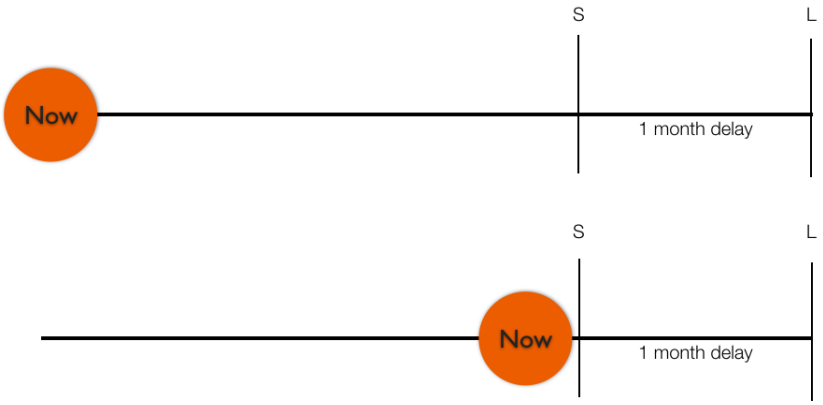

Figure 2: Consistency

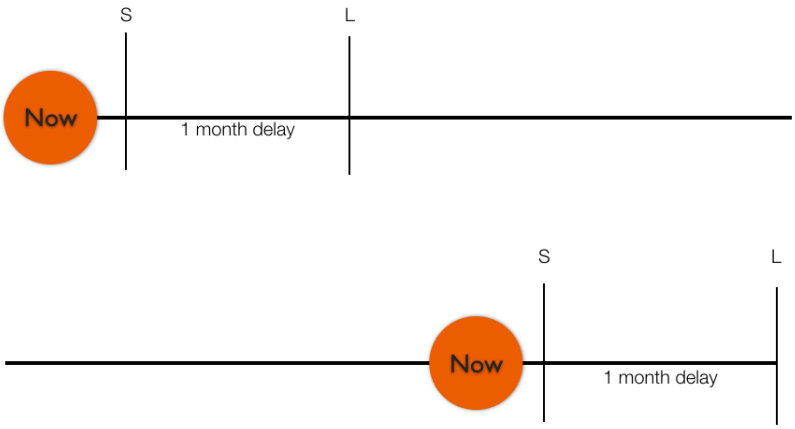

Figure 3: Invariance

The proof is trivial. As a result, we know that violating Stationarity while maintaining Invariance leads to violations of Consistency, and therefore, potential preference reversals.

What is Invariance? Continuing with the same terminology and constraints, a set of preferences is time invariant if

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

Invariance is the claim that "preferences are not a function of calendar time" (Halevy 2015, 341). Preferences are invariant under time translation. No moment of time has a special character or status. See Fig. 3.

The gap between Stationarity and Consistency is hugely important. In fact, it threatens to undermine the conventional wisdom surrounding most of our the empirical findings. Decades of experiments convincingly show that Stationarity is often violated. The paradigm for testing discounting elicits preferences at one time, not two. Logic tells us

that

$$\neg(\text{Stationarity}) \rightarrow \neg(\text{Consistency}) \text{ v } \neg(\text{Invariance})$$

So a violation of Stationarity, which is what is mainly tested, does not by itself mean that one is inconsistent (and therefore, irrational). Instead one may violate Invariance (or both Invariance and Consistency).

How bad is this gap between Stationarity and Consistency? Is Invariance merely a "technical" axiom, something that we can safely assume to get from non-Stationary preferences to preference reversals? Hardly. The few empirical studies that have been done testing not only Stationarity but all three (which necessitates testing subject at t and at a later t') reveal that many and perhaps most time non-Stationary subjects in fact *violate* Invariance and *not* Consistency (Halevy 2015; Janssens et al 2017).[6] With the admittedly limited evidence we have, non-expoential discounting can't be definitively connected to allegedly irrational preferences reversals at all.

If we want to connect violations of Stationarity to violations of Consistency, Invariance is the bridge we need. Stationarity is about preferences at one evaluation point. Consistency is about preferences at two or more evaluation points. They are logically distinct. Invariance allows us to link the two by shifting preferences at one time to those at another. Assuming Invariance, someone violating Stationarity is expected to honor their past preferences about that future date when it is present. In that way we can generate inconsistency and raise the specter of exploitation.

3.2 A New Master Argument for EDU

We need an argument for Invariance having normative force. To find one, we need only cast our gaze from social science to philosophy. Many phiosophers have independently defended a position that is essentially identical to Invariance. The consequentialist tradition figures prominently here. Treating different time stages of individuals and groups as akin to other people, many philosophers hold that we ought to show equal concern for each temporal stage, eschewing distinguishing any temporal stage–just as consequentialism is based upon equal concern for all individuals. Intertemporal tradeoffs are justified if they bring about greater lifetime utility, just as transfer of wealth from one person to another is justified if it brings about greater overall happiness. Here is Adam Smith 1790:

[6] In Janssens et al 2017 subjects in rural Nigeria were given choices between smaller-sooner and larger-later rewards. This was done at two times and designed to test all three time preferences. 43.4% of subjects violated Consistency. But only about half of these violated Consistency *and* Stationarity. The others violated Invariance. In fact, more violated Invariance than any of the other two conditions, showing that we definitely cannot infer inconsistency from a violation of Stationarity. Moreover, and interestingly, they found that "violating time consistency but not stationarity is correlated with reductions in wealth from the first to the second decision moment." Changing one's mind in the face of financial shocks, of course, might be the paragon of rationality, not a departure from it. These subjects suffered income decreases that they had not anticipated. (Of course this example can be interpreted in a way compatible with EDU being a good normative standard; I mention this example just to show that empirical departures from EDU may come from violating Invariance.)

> The impartial spectator does not feel himself worn out by the present labour of those whose conduct he surveys; nor does he feel himself solicited by the importunate calls of their present appetites. To him their present, and what is likely to be their future situation, are very nearly the same: he sees them nearly at the same distance, and is affected by them very nearly in the same manner. (VI.i.11)

And Henry Sidgwick 1874:

> My feelings next year should be just as important to me as my feelings next minute, if only I could be equally sure of what they will be. This impartial concern for all the temporal parts of one's conscious life is a prominent element in the common notion of the rational as opposed to the impulsive pursuit of pleasure.(1, 113)

But the thought is not at all restricted or necessarily tied to consequentialism. Here is John Rawls 1971:

> The mere difference of location in time, of something's being earlier or later, is not a rational ground for having more or less regard for it. (1971, 259)

The intuition is reasonably clear. Time, like space, is not a morally relevant property. Hence preferences that are based on what Parfit calls a "purely positional property" don't reflect an important intrinsic difference in what is desired. Lowry and Peterson 2011 call this the Standard Argument and Sullivan 2018 dubs it, as we shall, the Non-arbitrariness Argument. Our valuings should not be a function of purely positional properties.

Non-arbitrariness supports Invariance. Invariance – recall Fig. 3 – is the claim that one's preferences should survive a time-translation of the world forward along the time line. The philosopher Leibniz was worried about making arbitrary choices; he didn't want God to have to choose an arbitrary location for matter in Newton's spatially and temporally homogeneous arena. Instead of material contents, Invariance is the claim that our preferences are invariant under a shift in the time line. Non-arbitrariness offers reason to think they should be unchanged, for temporal position is a mere positional property lacking moral relevance.

Put together, we now have a kind of master argument for Stationarity, and from there, EDU (see Fig.4). Invariance is endowed with normative charge through the argument against arbitrariness. Consistency is endowed with moral charge via the threat of exploitation. Together they imply that we should satisfy Stationarity. Stationarity (plus Fishburn and Rubinstein's other axioms) gives us EDU, normatively charged. That's a lot of assumptions, true, but on the whole, it's progress. Whereas Stationarity on its own seems very hard to motivate, we've derived it from premises each of which have some claim to normative standing. We don't want self-sabotaging preferences that allow us to be exploited, nor do we want arbitrary preferences based on merely positional properties. Together, these desires lead to EDU.

That's the best argument I can muster for treating EDU normatively. In its favor, I can point out that both normative premises have independently been adopted elsewhere and that the path from these premises to EDU is a rigorous one.

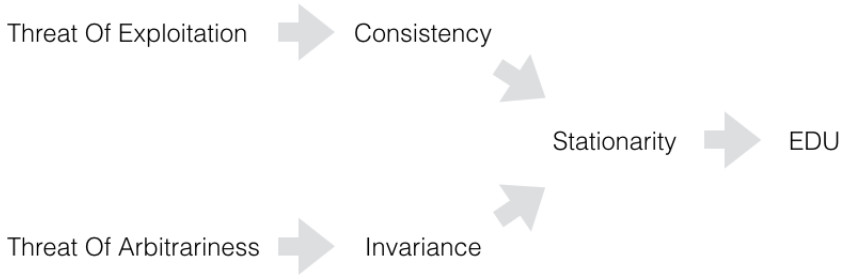

Figure 4: Master Argument

4 Initial Concerns

Before evaluating the main argument, I want to make a few points that challenge the overall narrative behind the Standard Model.

First, hyperbolic discounters are typically introduced as impulsive characters, people who are unwilling to make now-for-later sacrifices, unable to get up when alarm clocks sound, and so on. Ignore all of that moralizing and psychologizing. Hyperbolicity, of course, concerns only the form of the discounting function, not the content of the preferences nor the value of the discount rate. The exponential discounter can have preferences for gluttony, greed, exercise avoidance, and more. They may care only about the moment. The exponential discounter's only sure virtue, if it is one, is not deviating from *ex ante* plans. But those plans may be associated with any other epistemic or moral failing you like, plotting paths to sure ruin.

What drives much of the moralizing around hyperbolic discounters, I conjecture, is the association with Mischel's famous "marshmallow" tests (Mischel and Ebbesen 1970; Mischel, Ebbesen, and Zeiss 1972). In the delay-of-gratification paradigm children are offered a choice between two rewards, a large reward if the child waits for the experimenter to return, or a smaller reward if the child grows tired of waiting and gives up. Famously, ability to delay gratification is strongly associated with later achievements and socio-emotional behavior. These tests have recently come under fire because it seems socioeconomic class better predicts outcomes than these personality traits (Watts et al 2018). Even if this iconic result survives, it's important to appreciate that the delay-of-gratification paradigm is *not a test of discounting*. Subjects are told that they can have the preferred reward if they wait for the experimenter to return, but they are not told *when* the experimenter will do so. There is uncertainty in the timing of this return, uncertainty that is crucial to the very paradigm. In fact, McGuire and Kable 2013 make the case that in light of this uncertainty it is often rational for children to accept the smaller sooner reward. (Note, incidentally, the similarity to Jevons et al's diagnosis of temporal preference being the cause of poverty and the potential for the causal arrow between the two to flip.)

In general, hyperbolic functions are initially steeper and then later flatter than expo-

nential functions. I suspect that part of the moralizing about hyperbolic functions is a relic of the old neoclassical economic or philosophical outcry against steep discounting. That moralizing may be correct, but that is not relevant to the issue of exponential versus hyperbolic functions. The case for EDU is about the form of the function, not the values it adopts. And of course one can choose exponential or hyperbolic functions as steep or as flat as one likes.

Second, even if hyperbolic discounting were tied to steep discount rates, steep discounting also gets a bad rap that isn't always earned. Again we find a lot of unwarranted moralizing. A steep discounter is condemned for lacking many character virtues, such as being willing to make now-for-later sacrifices. We can all agree that now-for-later sacrifices are crucial features of many noble lives. Endurance athletes, good parents, hard workers, and more sacrifice untold hours of pleasure and opportunity for the satisfaction of a good race, good children, fine products. But utility-maximization equally demands later-for-now sacrifices. South Korea's long work week is widely considered a major problem. One suggested cause is a norm against later-for-now sacrifices. Postponing happiness too long is a very real problem. Trade-offs across a lifetime are necessary, and we shouldn't associate one way of doing this badly with the whole practice of discounting.

Third, that hyperbolic discounters *can* reverse preferences doesn't mean that they *will* in any finite period of time. The normative standard doesn't specify time periods or particular values of discount rate. A hyperbolic function can approximate an exponential one for as long as one wants, never getting to an actual reversal. In the toy example from above, our exponential discounter discounts according to the schedule $\{1, 0.5, 0.25\}$. Now consider a hyperbolic discounter using the schedule $\{1, 0.5, 0.24\}$. If the first two time periods are long, say, enough for a lifespan, there is no difference between the two. But also, even if one lives to the third time step, no bets exploiting the difference between 0.25 and 0.24 may actually be offered. Absent other information, a life without preference reversals can be modeled with an exponential or hyperbolic function.

Fourth, the evolutionary story associated with the Standard Model seems a bit shaky. Were our ancestors really offered so many clear choices between smaller-sooner rewards and later-larger rewards? How is this information transmitted? Biologically or through culture? Does \$100 next week really trigger an emotional response that \$110 a year and a week from now doesn't? Is this "hot" system really engaged and behind my 403b selections? I personally do not notice any arousal accompanying my investment decisions. Both my more prudent and less prudent investment decisions seem to require the same amount of cool attention/indifference and their performance takes the same amount of time. Also, EDU doesn't specify a particular discount rate ρ or unit of time t. The time t could be seconds, days, weeks, months, years. Are we to suppose that evolution doesn't care about that?

Fifth, exponential discounters *will* suffer a kind of preference reversal, contrary to common wisdom. Suppose you adopt a steep discount rate and burn through your fortune in a wild weekend. On Monday morning, waking up on the floor surrounded by empty bottles and unpaid bills, you regret your earlier choices. Few economists consider the extreme past-discounting that is familiar to us all: once a good has been consumed, it is "over and done" and we value it differently (Parfit 1984, Suhler and

Callender 2012). Arguably economists need to add this past discounting, for otherwise an exponential discount rate implies that we care more about temporally distant past outcomes than proximate past ones (Hedden 2015), which we don't. Economists who consider past discounting sometimes try to avoid this point by claiming that economics is future-oriented, concerned only with future actions, and then claim that this type of regret won't influence future-oriented actions. Depending on your level of risk aversion, however, that may not be true (see Dougherty 2011). Even exponential discounters, if modeled realistically as discounting only from the present moment onwards, can suffer preference reversals.

I don't wish to minimize the significance of these concerns. Some may be sufficient to warrant a rejection of EDU. Nonetheless I wish to add some further challenges that attack core parts of the Standard Model. Let's now return to the master argument built in Section 3.2. First I'll rehearse some reasonably well-known objections to linking rationality with immunity from exploitation. Then I'll make a more original argument that puts pressure on the distinction between pure and impure time preferences, thereby challenging Non-arbitrariness.

5 Consistency

In the master argument the path to Stationarity goes through two paths, one via the threat of exploitation motivating Consistency and the other via the threat of arbitrariness motivating Invariance. Let's focus on the first path. I want to challenge this path by first, reminding the reader of the many well-known problems with it, and two, by invoking the more direct problem of changing selves.

5.1 Is Susceptibility to Exploitation a Criterion of Irrationality?

We've assumed until now that susceptibility to exploitation is a symptom of irrationality. If one makes plans and knows in advance that they will leave one open to a loss of some kind, isn't that irrational? The answer is hardly clear. In the case of credences the status of these kind of dominance arguments has been debated for decades (see Vineberg 2016 and references therein). One common argument is that it's irrational to have credences that violate the axioms of probability theory because such credences open one up to Dutch books, i.e., a series of bets sure to lead to loss. Is such susceptibility really a mark of irrationality?

In reply, some point out that this assumes one accepts bets with ideally coherent bookies. The lesson instead might be that it's rational to not enter into bets with such ideal agents. The argument also assumes that mere potential for loss, as opposed to actual loss, is bad. There may not be bookies around able to make the bets necessary to take advantage of you. Or you may not think that there are–with good reason. Others highlight the gulf between what seems rational and what avoiding a Dutch book demands. Avoiding a Dutch book requires that you place credence in any logical truth at 0 or 1. But your evidence may have you uncertain about such a proposition and employing a credence of (say) 0.5 instead. This suggests a deep division between the type of rationality under

consideration and "internalist" conceptions of rationality where rationality is a matter of the evidence accessible to one (Foley 1993). Still others point out that there is a kind of mirror image of the Dutch book theorem that guarantees future gain as opposed to loss from incoherent degrees of belief. A way to break this symmetry is then needed.

Suffice to say the connection between Dutch books and incoherent degrees of belief is not rock solid.

Note, crucially, that all of the above reasons to resist the link in the credal case are also reasons for doubting the link in the present case of preferences. Consider the hyperbolic discounter considered above who changes their discount rates. Assume Invariance. She is therefore exploitable. But she is only exploitable by someone who knows that she will change his mind and the price points at which she will. We typically don't encounter such people, so it might be rational to assume we won't meet any. Perhaps also the world might work out for her such that the change of preference is beneficial to her in some way, all things considered. That there is a way the world might work out (say, betting with ideally rational agents) which is better if you don't change your mind is not enough for us to conclude that your preferences are irrational.

At the very least, we can say that the connection between exploitation and irrationality is very tenuous.[7] It's tenuous in the case of credences and at least as tenuous in the case of preferences.[8]

5.2 The Problem of Changing Preferences

Step back and think through a situation where a person satisfies all three conditions, Invariance, Stationarity and Consistency. When we see how demanding this is, I feel that it undermines the case for EDU being a normative standard. It also highlights the problem of changing preferences and its connection to Consistency.

Stationarity holds that, from the perspective at time t, a week is a week for your preferences, whether the week is next week or next year. Invariance slides these preferences at t to t', where $t' > t$, holding them unchanged. Consistency insists that there are no preference reversals at t'. All together, an EDU discounter is held at t' to their preferences about t' at t. That is, one is essentially demanding consistency between how they will feel in the future and how they feel now, even though those feelings about the future depend on what they think will happen and the actual future feelings will depend on what actually happens.[9]

Is that a fair requirement of rationality?

[7] See Moss 2014 for an excellent further discussion of this topic.

[8] In fact, Pettigrew 2018 argues that many defenses of the Dutch book argument in the credal case don't work with preferences.

[9] Note that this match is a bit like the one insisted upon by the controversial Principle of Reflection in Bayesian decision theory. Reflection states that if one's credence in a proposition A in the future, t', is r, then one's current credence at t in A should be r. That principle must be modified if it is to have any purchase, for counterexamples are easy to generate, e.g., if I know I will be drunk at t', then I should not treat future me as a guide to what my credences now should be. Here we are dealing with preferences, not credences, and further, we don't know what our future preferences will be. Like Reflection, however, EDU is demanding a kind of match across time.

No, not unless one adds many provisos. The problem is that we've run headlong into what has been called the *problem of changing selves* in decision theory (Pettigrew 2018). As time elapses from t to t', many things may change. Tastes change. New information arrives, e.g., learning. Old information departs, e.g., forgetting. One's underlying values or even self may change. For all of these reasons, when making a choice at t', it doesn't seem at all incumbent upon me to look back and match what I earlier preferred at t . If you asked someone why they bought a bright orange car and they replied that they did so because that is what they preferred ten years ago, that person might strike you as more strange than rational. If the earlier self guessed all of the intervening events between t and t' right, then yes, one might expect some match between preferences at those two times. Otherwise not. To demand a match puts too much weight on correcting anticipating what will happen and staying the same through time.

There are proposed solutions to the problem of changing selves. And there is motivation to finding one. After all, one wants to say that being able to make and stick to a plan is a sign of rationality. The problem is especially hard to solve when one considers values and selves (Pettigrew 2018). Both are constantly evolving. In the face of all this flux, is there anything we can insist on remaining invariant?

I can't possibly survey all the relevant moves and literature. What's clear is that we're not going to come through all of those moves with EDU intact as the normative standard. In each case we must retreat to something invariant, thus limiting the scope of EDU in problematic ways.

To get a sense of the uphill challenge, consider two natural responses. Hedden 2015 proposes (and later rejects) a principle dubbed Utility Conditionalization: "It is a requirement of rationality that your ultimate preferences –preferences over maximally specific possibilities–do not change over time." When I was young my favorite ice cream flavor was coffee. Now it's chocolate. Arguably this preference change is not a change of ultimate preference, for I still have the more specific preference for the flavor that I most enjoy when I would most enjoy it. That preference remains invariant. Similarly, subjects in Janssens et al 2017 may prefer to discount the future at one rate when financially well off and by another rate when not. In this way their ultimate preferences remain invariant too. Retreating to Utility Conditionalization means that ultimate preferences should be discounted exponentially. EDU would remain the normative standard only for them. Although EDU is proposed as a theory of ideal rationality in the presence of temporal discounting, this move to ultimate preferences would make the theory way too ideal for economics. The ultimate/non-ultimate divide is untestable. Whenever we meet non-exponential discounting we can always redescribe the case to find a constant discount preference lurking in Platonic heaven. Hedden catalogues many problems with this principle. But even apart from these I don't think this save of EDU is one social scientists will find useful.

Another response to the problem of changing selves is to let higher-order preferences decide who wins between two lower order preferences. Chocolate or coffee flavor? Coffee flavor seems to me more sophisticated, and I aspire to be sophisticated; hence, perhaps I should let this high-order preference break the tie between my preference for chocolate at t and for coffee at t'. But as Pettigrew 2018 argues, why do higher-order preferences

always "win" on this view? Often our lower-level preferences end up tutoring our higher-order ones in ways in which we approve. In fact, this solution lets us see a general problem challenging any solution to the problem of changing selves. We have preferences at t and then at t'. Does one trump the other? Or is there a third deeper or higher preference that trumps both? In general, we have a clash among preferences. If one is to have normative force, it seems that it must be a rational one. Matching my preference at t with that at t' isn't warranted if I know I'm going to be cognitively impaired at t'. We only want to demand cross-temporal preference matches with rational preferences, but this of course employs normative judgements not coming from EDU.

Finally, let me point out that a small but (to my perhaps biased eyes) increasing band of economists are similarly proposing so-called "time consistent hyperbolic discounting." What unites this group is a shared rationale for non-constant discounting. From the above theorem, we know this can be Consistent so long as it violates Invariance. Non-constant discounting violates Invariance because it makes preferences dependent on calendar time. The economists defending time consistent hyperbolic discounting are not motivated by the flux life presents. Rather, they are typically thinking about systematic reasons to be non-constant, and in particular, the underlying causes of discounting. They might, for instance, be thinking of a natural life cycle. I am a finite creature who will likely die in the next 35 years. I might take this into account and adopt a non-constant discount rate. Or they may be thinking about the effects of modeling anticipation for larger-later rewards. Like me, they are worried about insisting that one ought to be able to match one's expectation of future preferences with actual later preferences. Galperti and Strulovici 2014 write, "Koopmans' stationarity thus compares immediate 'real' consumption with anticipated one...By contrast, we adopt the view that anticipated consumption is radically different from actual consumption, a view inspired by the fact that the former consumption is purely imagined, while the latter has specific physiological and sensorial components." Put like this, it seems very demanding to insist on Consistency between these types of consumption.

Philosophers may want to cry "foul!" at this point, as introducing a natural life cycle or anticipation will introduce non-arbitrary time preferences. We'll get to that in a moment. Suffice to say, adding in realistic aspects of life or psychology only makes the rationality requirement of matching preferences that much more demanding.

6 Invariance

Invariance is motivated by the argument from non-arbitrariness. "Argument" is perhaps a bit strong, as the case really comes from the intuition that temporal position in and of itself is not a salient property to value. I want to put pressure on this idea. But first, let's work our way up to that discussion.

Everyone admits that that there are plenty of good reasons to discount future outcomes. Jevons, Smith, Sidgwick, Rawls and everyone else who has ever thought about the topic mention uncertainty. If the chances of obtaining a later reward aren't one, that will rightly affect my valuation of that future outcome. Rational actors take uncertainty

into consideration. Since the future is uncertain, and the further future generally more uncertain, time an be a proxy for uncertainty. What might look like temporal discounting can then really be perfectly legitimate discounting for uncertainty. The same is true of other features. You must wait for that larger-later reward. Will you "die of" anticipation for it? If so, you may wish to take account of this psychological state. You might with good reason want to discount or even inflate the value of that reward. The same is true if one takes into account pain of abstinence, differences in construal, optimism, intrapersonal empathy gap, personal identity, time perception, and more. Arguably some of these processes provide rational reasons to discount.[10]

The philosopher, however, is quick to dismiss all of these reasons as relevant with the invocation, "impure!" Non-arbitrariness holds only when considering so-called *pure time preferences*. Pure time preferences are preferences for a particular temporal position *independent of any non-temporal factor*. If we prefer the immediate reward because it is more certain or the anticipation will kill us, that is all fine, but then one is not discounting time but taking account of uncertainty and anticipation. Duration is simply associated with another non-temporal property that happens to occur in that same time interval. Mere temporal position, as opposed to duration, can also act as a proxy. Rawls advises that "...we should take into consideration how our situation and capacity for particular enjoyments will change" (1971, 93-94). Suppose one will enjoy a carnation only at one's high school dance at age 18. Then taking that calendar date into consideration when assessing the value of a carnation is again fine, for the date is a proxy for what one really cares about, the capacity to enjoy a carnation. Invariance is not threatened by the impure.

Restricting ourselves to pure time preferences, not everyone accepts the Non-arbitrariness intuition. Some view it as too strong if one considers Buridan's ass cases. In some situations one has no choice but to be arbitrary. You have tickets to go to the opera either this month or next month. All else is the same. You're inclined to go this month. Does having no rational ground for this preference mean that it is irrational, or is it instead rationally permitted? Arguably, the latter (Lowry and Peterson 2011; Żuradzki 2016). On this view, merely positional properties can be normatively neutral grounds for a *permissible* preference, contrary to Rawls and Parfit. While I accept this counter-argument, I want to go much further and put pressure on the distinction between pure and impure time preferences that underlies this discussion.

The pure versus impure distinction assumes that "mere" time passing doesn't alter the value of anything. That seems useful and clear, and often it is. However, as Ziff points out, "time, even in abstraction from all non-temporal considerations, still has a character" (Ziff 1990). Time has duration, order, and arguably, other features. Ziff emphasizes that this character can be pleasant, unpleasant or neutral. I want to add that this character can also be very hard to separate from the rest of the world, and when it is, often we are left with the preferences of some ideal observer who is outside of time, preferences that we may not view as our own. To get a sense of what I mean, let's go through some

[10] Ahmed 2018, using the distinctions employed in Section 4, argues that discounting for diminished personal identity is rational.

examples.

All human beings will age. Try as they will, Hollywood stars have not avoided it, and only fictional cases like Benjamin Button age in the other direction. We can distinguish at least three types of information related to aging: one's actual age, that aging happens, and that one has a typical finite lifespan. All three provide reasons to discount. When I was young I drove a car that was so faulty that I never filled up the gas tank all the way – I never expected the car to last that long. Knowing my age puts me in a similar position when considering the lifespan of purchases I make (roofs, solar panels, etc) and other major decisions (whether to have another child). Simply knowing that aging happens, even if not the actual age, also affects many preferences. I will discount the value of a toy, for instance, for I know that preferences change with age in various predictable ways. Knowing even that I am temporally finite, or better, a creature who will live under a hundred years as opposed to a thousand years, is reason to discount value at some future times. Being dead in the future is an excellent reason to discount.

Are these types of information pure? Presumably not. One's actual age reveals some information about one's capacity for particular enjoyments and aging reveals that these capacities will change. The same goes for lifespan, as that conveys information about a dramatic change in capacity.

Note how difficult it is to disentangle these features from time. Aging and the rest are connected to entropy increase which in turn is very intimately connected with the direction of time. True, thermodynamics and entropy increase and aging aren't logically implied by general relativity, our best theory of time. So they are logically detachable from time itself, like the carnation at dance case, and consequently they arguably warrant the impure classification. However, entropy increase and aging and their connection to time may well be nomological implications of the best package of laws of nature for our universe (Callender 2017). Possible worlds wherein macroscopic objects don't likely age in one direction along their worldline may be unphysical.

If this is right, then we can only tease apart some temporal from non-temporal features in worlds unlike ours. Discounting may be irrational in such worlds, but so what? The people and preferences implicated are those that don't associate temporal duration with any character at all (e.g., aging), don't know that life is finite, and more. Even if we admit that discounting in such "pure" worlds is irrational, we still face a serious question of why that is relevant in our world and to us. Compare this point to criticisms of ideal observer theories of the good. These criticisms complain that the process of idealization leave the ideal observer so unlike the actual person that it's hard to see how their good is your good. I'm pointing out that the process of "purifying" takes us to possible worlds that are so remote as to have little relevance to how you should actually discount.

Let's put the point another way. I concede that logically speaking we can separate the pure from impure. The idea is that our pure preferences should survive a time translation along the timeline. We pick up all the material contents of the world a la Leibniz and shift them forward in time. Apart from Buridan's ass type situations, that shouldn't matter to our preferences. However, that shift is essentially equivalent to *renaming* all

the moments of time.[11] Say we shift everything forward a year. Now "2019" is 2020, and so on. Arguably this redescription doesn't affect the value of anything. A good case for Invariance results. But this redescription is trivial. And as soon as we make it non-trivial, the good case vanishes. It vanishes because temporal durations always have characters and those characters matter to us in different ways.[12]

7 Conclusion

We've made as strong a case as we could for understanding EDU as the normative standard in future discounting. The threat of exploitation motivates a normative take on Consistency. The threat of arbitrariness motivates a normative take on Invariance. Consistency and Invariance together then led their normative weight to Stationarity, the crucial condition underlying EDU. Is this case compelling?

The defense perches blocks precariously upon one another in a high stakes game of Jenga. The blocks are delicately poised, swaying, ready to collapse at the slightest perturbation. I personally feel that the Jenga blocks collapse into shambles. Recall some of the major assumptions of the master argument:

- That being in principle susceptible to exploitation is sufficient for irrationality
- That we can solve the problem of changing selves (and solve it in a way compatible with EDU)
- That past discounting can be ignored
- That preferences for temporal patterns –e.g., not wanting fish two nights in a row – can be ignored
- That preferences for temporal positions aren't normatively neutral and permissible
- That we can cleanly distinguish pure and impure temporal preferences (without the purification making your preferences alien)

In addition, one could add some of the more basic assumptions, such as treating time preference as exogenous, utility as separable, and more, as well as the other axioms necessary to get from Stationarity to EDU.

Given our objections, I can imagine only one style of response: retreat to safer ground. EDU is an ideal model with many ideal assumptions. True, in many cases it does not fit real-world situations. But the closer the real world approximately matches the ideal assumptions, the more purchase EDU has normatively. For instance, consider setting an

[11] I write "essentially" because this claim ignores many technical issues about symmetries – especially regarding translations in curved spacetimes–that I feel safe in bracketing for current purposes.

[12] In fact, one can relate this problem to that of Section 5.2, as Richard Pettigrew helpfully mentioned to me. No matter how pure one goes, ultimately one can't eliminate the fact that Invariance refers to two different times, not one time. Hence one might violate the condition merely by changing one's mind about how to discount as opposed to picking out any temporal location as special. Again see Moss 2014.

alarm at night to wake up for work the next day. We can be reasonably confidant that you will be roughly the same person the next morning, that you still want an income, that the eight hours of aging won't matter, that your being eight hours closer to death isn't a concern, and so on and so forth. In that case – and in many others where similar concerns can be put aside – it makes sense to not be arbitrary and to not adopt a potentially self-sabotaging set of preferences. In those circumstances EDU is the normatively correct policy for time preferences.

Fair enough. I personally see more flux than invariance in our preferences and circumstances, so I suspect these situations where EDU rules are rare and insignificant. I note also that this response doesn't answer all our worries, e.g., about past discounting. Here let me note that even if one retreats along these lines, or similar ones, none of this in any way saves the Standard Model widely employed throughout social science. Social scientists should be very uncomfortable with this line of defense. The Standard Model states that we're known to depart from the rational standard. But that's hardly true if this is right. Behavioral economists and psychologists have a hard enough time controlling for confounds such as gender, race, age, and socioeconomic status. Now they must control for *literally everything non-temporal*. They haven't, so we don't know if we're being irrational. Nor can they – or even come close – so we'll never be able to say whether we violate this normative standard. The same goes if we retreat to EDU operating only on ultimate preferences. The ultimate/non-ultimate division is untestable, so again the empirical studies can not be interpreted as they have been.

EDU is a wonderful tool – for instance, when working through decisions involving compound interest – but I don't think it is normatively compelling in general for all intertemporal decision making. Furthermore, the Standard Model surely has many correct components to it. A lot of good science lies behind it. With the normative foundation removed, however, we may have to re-interpret much of this science.

The primary object of re-interpretation, I suspect, is the idea that "temporal anomalies" are departures from rationality. These preference patterns may not be problematic. That remains to be seen. Showing that they do not conform to EDU is not sufficient for the charge of irrationality. If correct, this can have important consequences.

For example, today researchers throughout the social sciences try to connect hyperbolic discounting and other "anomalies" of intertemporal decision-making to negative social and personal outcomes. Obesity, drug addiction, failure to save, and much more, are all connected to hyperbolic discounting. Some studies even suggest that one's time preference is a stable personality trait, possibly even heritable and genetic in origin. Hyperbolic time preferences, because tied to irrationality, are automatically seen as a cause of these problems, just as in neoclassical economics. Since there are various possible ways of manipulating our time preferences – through behavioral, neuromodulatory, and pharmaceutical means – this setup suggests possible interventions.

For example, some evidence links our time preferences with (say) addictive disorders. Other evidence suggests, as mentioned, that our time preference is hereditable and partly genetic in origin. If time preference risk alleles for drug abuse are found, one can imagine a policy of genetic pre-screening for these alleles. As a means of prevention, patients with a high risk profile could then undergo time preference therapy to decrease their chances

of drug addiction (see Gray and MacKillop 2015 for a critical evaluation of this idea).

If EDU is rejected as a normative standard, then all of this may be the wrong way around. We may have to admit, as Irving Fisher begrudgingly did, that this time preference can be rational. Pills to fix unhealthy time preferences would have the causal arrow wrong. Indeed, there is a lot of work suggesting that our time preferences are rational. A steep hyperbolic discount rate may be the height of rationality for people unfortunate enough to have serious hazards and uncertainty scattered throughout their future. If mortality is high, meals scarce, safe lodging uncertain, and so on, then it may be that hyperbolic discounting maximizes utility (Burness 1976; Farmer et al 2009; Frankenhuis et al 2016; Griskevicius et al 2010; Pepper and Nettle 2017; Sozou 1998). Rather than blame people's traits for failing a dubious standard taken out of context, this direction of the causal arrow forces us to concentrate on the question of why people are in the conditions such that steep hyperbolic discounting is rational. Different policy interventions are suggested as a result.

If EDU is rejected as the normative benchmark, what should replace it? Developing a full answer requires another paper, but I suggest that spatial discounting provides an important clue. As with time, we discount for spatial distance and order. Spatial discounting is studied in questions asking how far people prefer to live from dumps, trailheads, and so on. Unlike time, no one pulls space out of utility and treats it as an exogenous parameter. No one takes spatial discounting to be driven by a personality trait, stable disposition, or least of all, a hereditable property.[13] No one speaks of "pure" spatial discounting. It's all impure: roads, bridges, noises, smells, and more. As a result, what we regard as rational when discounting spatially is deeply contextual. We would not be surprised, therefore, to learn of dozens of "anomalies," dozens of possible causes of empirical patterns, and so on–some rational, some not. We would expect them. We would expect, in other words, just what we find in the temporal case (Urminsky and Zauberman 2016), suggesting that the two are more alike than is traditionally thought.[14]

References

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

[2] Anokhin A.P., et al. 2014. The Genetics of Impulsivity: Evidence for the Heritability of Delay Discounting. Biol. Psychiatry 77:887–894.

[3] Bentham, J. 1970, An Introduction to the Principles of Morals and Legislation, ed. J. H. Burns and H. L. A. Hart.

[4] Bickel W.K., et al. 2014. The Behavioral- and Neuro-economic Process of Temporal Discounting: A Candidate Behavioral Marker of Addiction. Neuropharmacology 76:518–27.

[13] For reasons to think time preference is malleable, see Lempert and Phelps 2016.

[14] Many thanks to Arif Ahmed, Jonathan Anomaly, Nancy Cartwright, Jonathan Cohen, John Dougherty, Tom Dougherty, Richard Pettigrew, and Samuel Rickless.

[5] Bohm-Bawerk, E. von. 1890. [1884]. Capital and Interest: A Critical History of Economical Theory. London: Macmillan Company.

[6] Bommier, A. 2008. Rational Impatience?, HAL-SHS hal-00441880, CNRS.

[7] Brink, D. 2011. Prospectus for Temporal Neutrality. In Callender C. The Oxford Handbook of Philosophy of Time (pp. 353-381). Oxford: Oxford University Press.

[8] 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(3), 547-549.

[9] Callender, Craig. 2017. What Makes Time Special? Oxford University Press.

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

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

[12] Dougherty, T. 2011. On Whether to Prefer Pain to Pass, Ethics 121, 521-537.

[13] Drouhin, N. 2009. Hyperbolic discounting may be time consistent. Economics Bulletin 29, 2552–2558.

[14] 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.

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

[16] Fisher, I. 1930. The Theory of Interest. Macmillan New York.1930, 73; see also 1907, 94.

[17] Foley, R. 1993. Working Without a Net: A Study of Egocentric Epistemology. NY: Oxford University Press, 1993.

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

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

[20] Gray, J. and MacKillop, J. 2015. Impulsive Delayed Reward Discounting as a Genetically-Influenced Target for Drug Abuse Prevention: A Critical Evaluation. Frontiers in Psychology 6: 1104.

[21] Griskevicius V, Delton AW, Robertson TE, Tybur JM. 2010. Environmental Contingency in Life History Strategies: the Influence of Mortality and Socioeconomic Status on Reproductive Timing. J. Pers. Soc. Psychol. 100, 241–254.

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

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

[24] 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.

[25] Jevons, W.S. 1871. Theory of Political Economy [1911], 4th ed., ed. H.S. Jevons, London: Macmillan.

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

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

[28] Lempert, K. and Phelps, E. 2016. The Malleability of Intertemporal Choice. Trends Cogn Sci. 20, 64–74.

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

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

[31] Loewenstein, G., and Elster, J. 1992. Choice over Time. NY, NY. Russell Sage Foundation

[32] Loewenstein, G., and O'Donoghue, T. 2005. Animal Spirits: Affective and Deliberative Processes in Economic Behavior. Working Paper 04-14, Center for Analytic Economics, Cornell University.

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

[34] Marshall, A. 1883. The Housing of the London Poor, Contemporary Review 45, 224–31; 1890 Principles of Economics [1930], 8th ed. London: Macmillan.

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

[36] Metcalfe, J., and Mischel, W. 1999. A Hot/Cool-System Analysis of Delay of Gratification: Dynamics of Willpower. Psychological Review, 106, 3–19.

[37] Mischel, W., and Ebbesen, E. 1970. Attention in Delay of Gratification. Journal of Personality and Social Psychology 16(2), 329-337.

[38] Mischel, W., Ebbesen E.B., and Zeiss, A.R. 1972. Cognitive and Attentional Mechanisms in Delay of Gratification. Journal of Personality and Social Psychology 21, 204-218.

[39] Moss, S. 2014. Credal Dilemmas. Noûs 48, 665-683.

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

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

[42] Pepper, G. and Nettle, D. 2017. The Behavioural Constellation of Deprivation: Causes and Consequences. Behavioral and Brain Sciences

[43] Pettigrew, R. 2018. Choosing for Changing Selves. Forthcoming.

[44] Pigou, A.C. 1903. Some Remarks on Utility, Economic Journal, March, 58–68

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

[46] Rawls, J. 1971. A Theory of Justice, Cambridge, Massachusetts, Harvard University Press.

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

[48] Sidgwick, H. 1874. The Methods of Ethics, London, MacMillan.

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

[50] Sozou, P.D., 1998. On Hyperbolic Discounting and Uncertain Hazard Rates. Proc Biol Sci. 265, 2015–2020.

[51] Story, G. W., Vlaev, I., Seymour, B., Darzi, A., & Dolan, R. J. 2014. Does Temporal Discounting Explain Unhealthy Behavior? A Systematic Review and Reinforcement Learning Perspective. Frontiers in Behavioral Neuroscience, 8, 76.

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

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

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

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

[56] 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.

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

[58] 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/>

[59] Watts, T.W., Duncan, G.J., & Quan, H. 2018. Revisiting the Marshmallow Test: A Conceptual Replication Investigating Links Between Early Delay of Gratification and Later Outcomes. Psychological Science 1-19.

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

[61] 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.