In Search of Lost Nudges

Guilhem Lecouteux

École Polytechnique, Department of Economics, Palaiseau, France

Published in *Review of Philosophy and Psychology* (2015), 6(3): 397-408. The final publication is available at http://link.springer.com/article/10.1007/s13164-015-0265-0

Abstract: this paper discusses the validity of nudges to tackle time-inconsistent behaviours. I show that libertarian paternalism is grounded on a peculiar model of personal identity, and that the argument according to which nudges may improve one's self-assessed well-being can be seriously questioned. I show that time inconsistencies do not necessarily reveal that the decision maker is irrational: they can also be the result of discounting over the degree of psychological connectedness between our successive selves rather than over time (Parfit 1984, *Reasons and Persons*, Oxford University Press). Time inconsistency can call for paternalism if and only if we accept that an individual is characterised by stable "true" preferences over time-dependent outcomes, and that she is rationally required to make time-consistent choices. This model is descriptively and normatively questionable. I then argue that behavioural findings may still justify paternalistic interventions, but on a non-welfarist basis.

Keywords: nudges, libertarian paternalism, Parfit, hyperbolic discounting, identity, psychological connectedness.

Acknowledgements: I am grateful to the editors and three anonymous reviewers for their comments and suggestions that substantially improved this paper. I also thank Robert Sugden for his careful rereading of the paper.

"We do not succeed in changing things according to our desire, but gradually our desire changes. The situation that we hoped to change because it was intolerable becomes unimportant. We have not managed to surmount the obstacle, as we were absolutely determined to do, but life has taken us round it, led us past it, and then if we turn round to gaze at the remote past, we can barely catch sight of it, so imperceptible has it become." (Proust (1927), *In Search of Lost Time*, vol. 6, *The Sweet Cheat Gone*)

1. Introduction

Standard economic theory is built on the assumption that people act as if seeking to satisfy stable and coherent preferences, and are instrumentally rational given their beliefs and those preferences. Behavioural economics however provides evidence that real individuals are subject to numerous decision biases, such as optimism and overconfidence (Sunstein 1998), loss aversion (Kahneman and Tversky 1979) or status quo biases (Samuelson and Zeckhauser 1988). The most common interpretation of behavioural findings is that real individuals – unlike rational *Homo economicus* – are not good at decision making, and often make *mistakes*: they would in fact choose differently had they "complete information, unlimited cognitive abilities, and no lack of self-control" (Sunstein and Thaler 2003, p.1162). This observation is the starting point of a general argument against Mill's Harm Principle (1859), according to which "the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others. His own good, either physical or moral, is not sufficient warrant". Conly (2013) argues for instance that our poor instrumental reasoning prevents us from achieving our own goals: it is therefore in our interest to accept that the government may prevent us from acting in accordance with our decisions.

The proponents of *libertarian paternalism* (Sunstein and Thaler 2003, Thaler and Suntein 2008, henceforth "TS") defend this paternalistic position by arguing that paternalism is actually inevitable. People are highly sensitive to framing effects (Kahneman and Tversky 2000), and their decisions are influenced by features of the choice environment that are not relevant from the perspective of

the social planner (Bernheim and Rangel 2009, p.55). Since there is no "neutral" frame, any way of presenting a situation of choice will necessarily make an option more salient than another one (by defining for instance a default option, a first option in a list of alternatives, etc.). TS argue then that the individual in charge of the design of the choice environment – the *choice architect* – should choose the choice architecture so as to help people to improve their well-being, "as judged by themselves" (TS, p.5). The logic of libertarian paternalism (henceforth "LP") is therefore that people aim to choose the options that will make them better off, but that – due to human fallibility – they often make non-rational choices, and miss their objectives. Since the choice architect has the possibility of slightly influencing people's choices, she should *nudge* them so that they achieve *in fine* their goals. A "nudge" is then defined as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any option or significantly changing their economic incentives" (TS, p.6). Since the individuals are not forced to choose the option the choice architect wants them to choose, Sunstein and Thaler argue that nudges preserve individuals' freedom of choice.

As an illustration, imagine the savings choices of an optimistic individual, Oscar. When hired for his first job, Oscar had to decide how much he wanted to save for his retirement. He was then convinced that he would quickly get a higher salary, and therefore – so as to smooth his consumption over time – preferred to consume a relatively large proportion of his income at the beginning of his career. He knew that this strategy could be risky, and therefore planned to decrease his level of consumption in the future if he could not manage to get a better-paid job. Unfortunately for him, Oscar never earned a significantly higher salary, and – due to inertia in his consumption habits – was also unable to keep his initial commitment. He then ended up with a quite low old age pension. Oscar now regrets his past decision and thinks that, were he able to turn back the clock, he would change his choice, based on better information about his ability to keep commitments and not misled any more by his optimistic expectations. Behavioural economists know that many individuals, like Oscar, are too optimistic (concerning their beliefs about getting a higher salary, but also about keeping their commitments) and are probably making a mistake when saving little for their

retirement. We should therefore nudge them today such that they will benefit from higher savings when retired. This can be done by exploiting their bounded rationality through framing effects: we could for instance create a default option such that they will be inclined to save more without significantly limiting their freedom of choice (Thaler and Benartzi 2004).

The two central claims of LP are therefore that (i) we can make the individuals better off, "as judged by themselves", since they can suffer from making mistakes, and (ii) we can do so while preserving their freedom of choice. The first claim characterises the paternalistic dimension of nudges, while the second claim ensures that nudges are compatible with libertarian principles. Several authors however question the claim that nudges are actually paternalistic, such as Hausman and Welch (2010, p.136), who argue that most of the nudges defended by TS are cases of rational persuasion. Since the choice architect is not supposed to impose her own normative views on the individuals, and since individuals are not constrained by the choice architect, they claim that TS's definition of "paternalism" is unsatisfactory – LP would be better understood as *beneficence*. Thaler and Sunstein (2003, p.175) indeed define a policy as paternalistic if "it is selected with the goal of influencing the choices of affected parties in a way that will make those parties better off": LP is therefore a *means* paternalism rather than an *ends* paternalism (Sunstein 2014, pp.19-20). Unlike ends paternalists who pursue objectives that are different from the ends of the individuals, means paternalists want to help people to achieve their own ends.

The object of this paper is to highlight that, within the context of intertemporal choices, LP implicitly relies on an implausible model of identity, and that referring to a more complex and plausible model may seriously undermine its paternalistic claim (according to which nudges can improve our *own* well-being). I firstly highlight that imposing a nudge on young-Oscar on the basis that retired-Oscar would have regretted his past choice requires accepting that (1) Oscar's true preferences remain stable over time, (2) retired-Oscar retrospectively knows what was in the interest of young-Oscar, and (3) Oscar is rationally required to use a constant discount factor when comparing future utilities (section 2). I then suggest analysing Oscar's choice in terms of Parfit's notion of psychological

connectedness, and show that the regrets of retired-Oscar cannot justify unambiguously a paternalistic intervention on young-Oscar (section 3). I finally discuss the possibility of justifying paternalism on a non-welfarist basis (section 4).

2. Regrets and mistakes

Libertarian paternalists argue that the choice architect should nudge young-Oscar today, because young-Oscar is making a mistake when saving only a small proportion of his income. We can legitimately assume that retired-Oscar is likely to regret the choice of young-Oscar because, today, many individuals are in the situation of retired-Oscar, i.e. they did not save a lot for their retirement and now regret their past savings choices. "Oscar" should therefore be seen as a statistically representative individual[1]. A nudge in this situation can be seen as a form of means paternalism if and only if young-Oscar would agree with the policy if he knew that he is likely to regret his choice later. I show in this section that this condition, although quite intuitive, is true if and only if we accept a rather implausible model of individual identity.

The economic analysis of intertemporal choices assumes that the individual "Oscar" is a set of transient selves representing a decision maker at different dates. For sake of clarity, I will denote by *t*-Oscar the self of Oscar at date *t*. Each *t*-Oscar has preferences over time-dependent outcomes[2] (x, n) – the promise to receive the outcome x at date $(t+n)$. LP is grounded on the idea that *t*-Oscar can be characterised by two different types of preferences: his *true preferences* – the counterfactual preferences on which he would act if he had "complete information, unlimited cognitive abilities, and no lack of self-control",

[1] We can notice that a few individuals may suffer from the nudge (those who truly prefer to consume a lot today): the legitimacy of nudges may therefore be questioned in this situation, since nothing justifies *a priori* that such exceptional individuals can be sacrificed to the benefit of the greatest number. See Bovens (2009, p.211) on this point.

[2] Throughout the rest of the paper, I will simply say "*t*-Oscar preferences" instead of "*t*-Oscar preferences over time-dependent outcomes".

whose associated utility function determines *t*-Oscar's welfare – and his *revealed preferences* – the preferences that are revealed by his choices (or equivalently, the preferences that determine his choices). LP's paternalistic claim relies on the assumption that there exists a discrepancy between young-Oscar's true and revealed preferences: young-Oscar should therefore be nudged such that he satisfies *in fine* his true preferences. Retired-Oscar's regrets are then taken as an indicator of the mistake of young-Oscar.

However, if retired-Oscar regrets the choice of young-Oscar, then it means that if *young*-Oscar had chosen differently, *retired*-Oscar would be better off. But under which conditions do we know that improving the well-being of retired-Oscar is actually in the interest of young-Oscar (and therefore that nudging young-Oscar on the basis that it will benefit retired-Oscar is actually in young-Oscar's interest)? We suggest that accepting the claim that retired-Oscar's regrets are a sufficient reason to nudge young-Oscar requires accepting the three following hypotheses:

(1) all the *t*-Oscars have the same true preferences,
(2) retired-Oscar's revealed preferences correspond to young-Oscar's true preferences,
(3) each self *t*-Oscar is rationally required to make time-consistent choices.

Those conditions (and the model of individual identity that supports them) are implicit in LP, and are not explicitly endorsed by its proponents: I nevertheless suggest that those conditions are necessary for LP's paternalistic claim, although they are both descriptively and normatively questionable.

If condition (1) is not verified, then time inconsistency does not matter from the perspective of young-Oscar: the reason why retired-Oscar regrets young-Oscar's choice is simply that his true preferences changed over time – and not that young-Oscar's choice was irrational. Indeed, allowing for preference changes with ageing may justify that retired-Oscar's regrets are consistent with a rational

choice of young-Oscar. The paternalistic claim therefore requires the stability of *t*-Oscar's true preferences over time[3].

Suppose therefore that condition (1) is verified. The reason why retired-Oscar disagrees with young-Oscar's choice is that they do not have the same *revealed* preferences, although they share the same *true* preferences. However, so as to be sure that young-Oscar would benefit from the nudge suggested by retired-Oscar, we need to assume that retired-Oscar's revealed preferences correspond to young-Oscar's true preferences (condition (2)): we should therefore assume that young-Oscar is mistaken, while retired-Oscar has a correct *ex post* assessment of young-Oscar's choice. We can for instance consider that young-Oscar's perception of his own interest is biased by a *present bias* (O'Donoghue and Rabin 1999), and has a tendency to put relatively higher weights on immediate outcomes – this would be the reason why he chose to postpone his savings effort. *Hyperbolic discounting* – the tendency to care relatively less about the outcomes distant in time, as involved by a present bias – is indeed a well-documented phenomenon (Frederick *et al.* 2002), and would explain young-Oscar's time inconsistent choice.

The difficulty of this argument is that nothing justifies *a priori* the rationality of retired-Oscar – in particular if we assume that young-Oscar presents a present bias. It is therefore not certain that retired-Oscar's revealed preferences actually correspond to his true preferences (and therefore, by condition (1), to young-Oscar's true preferences). Retired-Oscar may for instance present a *bias towards the future* (Parfit 1984, p.165). Suppose for instance that the satisfaction of young-Oscar's true preferences actually implied saving little for his retirement. When retired-Oscar is reminded of the pleasant life he had when he was young, he should accept that his low pension is the legitimate cost for his past consumption. But if retired-Oscar is biased towards the future, then, when comparing the expectation of a future consumption of €100 with the memory of a past

[3] I will discuss in section 4 the possibility to justify paternalistic interventions if condition (1) is not verified, since the choice of *t*-Oscar can be seen as a choice involving several individuals (with different true preferences). This will however not be a case of means paternalism, and will not support the paternalistic claim of LP.

consumption of €100, he will prefer the future consumption of €100. What retired-Oscar really wants in this situation is to go back 40 years earlier to change his decisions, and then immediately enjoy the long term benefits of his choice 40 years later: the regrets he expresses today do not necessarily mean that his past consumption was a mistake, since the cost supported by young-Oscar is almost imperceptible from retired-Oscar's perspective. The effort that young-Oscar perceived as "intolerable" became retrospectively "unimportant" for retired-Oscar. In this situation, saying that retired-Oscar would have agreed with being nudged when he was young does not mean that it would have been true for young-Oscar. Nothing therefore justifies *a priori* the empirical validity of condition (2).

Suppose now that conditions (1) and (2) are descriptively accurate. The last condition implicitly stated by LP to ensure that retired-Oscar's regrets reveal the irrationality of young-Oscar's choice is that young-Oscar is rationally required to make time-consistent choices, i.e. to make choices with which all his future selves would agree. Unlike conditions (1) and (2), condition (3) is a normative rather than descriptive condition: it indeed states how *t*-Oscar is rationally required to discount his future utilities, under TS's conditions of unlimited cognitive abilities, perfect information and complete self-control. LP is indeed a "prescriptive approach", i.e. it is an "[attempt] to offer advice on how people can improve their decision making and get closer to the *normative ideal*" (Thaler and Benartzi 2004, p.S167, our emphasis). The *exponential discounting model* (Samuelson 1937), which forbids time-inconsistent behaviours, is generally considered as a relevant normative model of intertemporal choices (and is implicitly considered as such by LP, since time-inconsistent choices are not allowed). This claim is for instance defended by O'Donoghue and Rabin (1999), according to whom time inconsistency leads to important welfare losses, and therefore that an exponential discounting may be preferable in terms of welfare (welfare is defined within their framework in a "long-run perspective" (p.113), with an equal weighing of the utilities of each period). Furthermore, we can notice that, if condition (1) is true, then, under TS's conditions of perfect rationality and complete information, (3) is necessarily true. If t_0-Oscar is rational (in the sense that he ought to choose his action so as to satisfy his true preferences), then he should take the exact same

decision when comparing two outcomes in t_0 and t_1, or in t_n and t_{n+1}. He indeed knows that, in n periods, he will face the situation of choice he faces today in the first case. If we assume that Oscar prefers €100 immediately to €100 in the future, then we can define a discount factor $\delta_{0,0} < 1$ such that t_0-Oscar is indifferent between €100* $\delta_{0,0}$ in t_0 and €100 in t_1 (a discount factor $\delta_{t,n}$ should be read as the discount factor used by t-Oscar when comparing two outcomes at dates $(t+n)$ and $(t+n+1)$). Hyperbolic discounting means that t_0-Oscar tends to give a relatively higher value to immediate rewards or costs than more distant ones: the discount factor $\delta_{0,n}$ between two dates t_n and t_{n+1} therefore increases when n increases, i.e. when the delay between the choice and its realisation increases. t_0-Oscar therefore chooses his level of savings as if he believed that he would be more patient in the future than he is today: but since Oscar's true preferences remain stable over time (condition (1)), then retired-Oscar will regret the choice of young-Oscar, since he is exactly as patient as were young-Oscar. Hyperbolic discounting therefore implies that young-Oscar will take decisions that retired-Oscar will regret: this therefore justifies a paternalistic intervention on young-Oscar in his own interest.

I have shown that the argument according to which retired-Oscar's regrets are a sufficient reason to nudge young-Oscar means that (1) those regrets are meaningful for young-Oscar, and therefore retired-Oscar and young-Oscar have the same true preferences, (2) retired-Oscar retrospectively knows what was in young-Oscar's best interest, and (3) young-Oscar should have used a constant discount factor when choosing his level of savings. Accepting LP's argument therefore implies that we should also accept the idea that Oscar is simply a set of transient selves t-Oscar with constant true preferences over time. This picture of Oscar's identity seems quite implausible, since it imposes the stability of individual true preferences over time-dependent outcomes (from a descriptive perspective), as well as the superiority of the exponential discounting model (from a normative perspective). I will now show that an alternative account of Oscar's identity does not necessarily support the normativity of the exponential discounting model, and that it is possible to rationalise time inconsistent behaviours (nudging young-Oscar would therefore not be a case of means paternalism any more).

3. Does time inconsistency matter?

The object of this section is to investigate whether *t*-Oscar is rationally required to discount his future utilities with a constant discount factor or not. We should firstly notice that, although people actually discount future utilities, it is not clear whether they are rationally required to do so (and reciprocally, whether discounting one's future utilities is rational or not). Indeed, if we assume that Oscar's objective is that his life goes as well as possible, as a whole, then it is not certain that there is any decisive argument for discounting future utilities (see for instance Broome (1991)). Frederick (2003) notes that temporal neutrality (the claim that a person should give the same weight to all utilities, regardless of their temporal position) implicitly assumes "that all parts of one's future are equally parts of oneself; that there is a single, enduring, irreducible entity to whom all future utility can be ascribed" (p.90). If we accept the existence of such an irreducible entity, then *t*-Oscar should not discount his future utilities.

By opposition to this "simple" view of identity, Parfit (1984), among others, offers a "complex" view, according to which such an irreducible entity does not exist: a person is a sequence of overlapping selves who are connected by different physical and psychological properties. Parfit suggests that there exists a relation of "psychological continuity" (p.206) between t_0-Oscar and t_n-Oscar: there exist strong psychological connections between t_0-Oscar and t_1-Oscar (such as shared memories, values, beliefs, desires...), as well as between t_1-Oscar and t_2-Oscar, until t_{n-1}-Oscar and t_n-Oscar. The relation of "strong psychological connectedness" is however not necessarily transitive: although Oscar can consider himself as the same individual than the individual he was yesterday, and that the one he will be tomorrow, he may not consider himself as the same individual than the one he was 10 years ago, nor than the one he will be in 10 years. Although there is some physical and psychological continuity between t_{-10}-Oscar, t_0-Oscar, and t_{10}-Oscar, they are not necessarily the same person, because they do not necessarily share the same memories, values or preferences.

Parfit then argues that, from the standpoint of the decision maker, it is not personal identity but psychological connectedness that matters (Parfit 1984, p.245): knowing that your personality is likely to evolve over time, you are not rationally required to care as much about your further future than your closer one (p.158). If we accept the complex view, then the different selves of Oscar can be seen as different persons: it is therefore not irrational for *t*-Oscar to discount his future utilities, since his future selves are not entirely himself. The idea that my future selves are not "entirely" myself, although I may have a lot of common with them, can be captured by the notion of *psychological distance* between temporal selves. Psychological distance relates to the difficulty for people to experience the feelings and subjective states of others (either their future selves, Wilson and Gilbert (2003), or other individuals, Andersen and Ross (1984)). Liberman *et al.* (2007) define for instance "psychologically distant things […] are those that are not present in the direct experience of reality" (p.353): at date t_0, since t_1-Oscar does not exist yet, t_0-Oscar cannot directly experience t_1-Oscar's utility. If t_0-Oscar is unable to fully experience the utilities of his future selves, then he is legitimated to discount their utilities. The existence of a non-null psychological distance between me and my future selves may therefore rationally justify the discounting of my future utilities.

The question that follows is therefore the determination of a criterion to measure this psychological distance between temporal selves. Such a criterion may then give insights into how the individual is rationally required to discount future utilities. If we show that there exists a plausible measure of psychological distance such that hyperbolic discounting is not irrational, then the paternalistic claim of LP will be seriously undermined (the exponential discounting model would indeed not be the unique defensible normative model of intertemporal choice). Since the notion of psychological distance is related to the ability to experience the utilities of one's future selves, I suggest defining the psychological distance $0 \leq d_{0,n} \leq 1$ between t_0-Oscar and t_n-Oscar as the loss of welfare experienced by t_0-Oscar when he gives €1 to t_n-Oscar. The distance $d_{0,n}$ therefore captures the idea that t_0-Oscar is less able to experience the utility of t_n-Oscar than his own utility.

We can now redefine the discount factor $\delta_{0,n}$ in terms of psychological distance. Recall that t_0-Oscar is indifferent between €100 in t_{n+1} and €100*$\delta_{0,n}$ in t_n. The difference between those two outcomes can be interpreted as the cost supported by t_0-Oscar when t_n-Oscar gives €100 to t_{n+1}-Oscar. This cost can therefore simply be measured as the difference between what t_0-Oscar would experience if he gave €100 to t_n-Oscar – i.e. €100* (1-$d_{0,n}$) by definition of the psychological distance – and what he would experience if he gave €100 to t_{n+1}-Oscar – i.e. €100* (1-$d_{0,n+1}$). We have therefore:

$$(1 - \delta_{0,n})*100 = (1 - d_{0,n})*100 - (1 - d_{0,n+1})*100$$

$$\delta_{0,n} = 1 - (d_{0,n+1} - d_{0,n})$$

From the perspective of t_0-Oscar, if two temporal selves are relatively close, then the discount factor $\delta_{0,n}$ tends to 1 (t_0-Oscar therefore assigns similar weights to selves who are relatively close, from his perspective).

Consider firstly LP's description of Oscar's identity. So as to ensure that condition (3) is verified, we must have a measure of the psychological distance between two successive selves such that ($d_{0,n+1} - d_{0,n}$) does not depend on n (otherwise Oscar is likely to make time-inconsistent choices). Furthermore, we can notice that, if t-Oscar is rational and does not hold false beliefs, then he *knows* that condition (1) is true. He could therefore perfectly anticipate the experience of his future selves: the psychological distance between the different selves t-Oscar is therefore necessarily null, implying that $\delta_{t,n} = 1$, for all t and n. LP's conditions therefore imply that t-Oscar is rationally required to be temporally neutral[4].

[4] A solution would be to consider t_0-Oscar's probability of dying before t_n: since there is a probability that t_n-Oscar does not exist, t_0-Oscar would then be able to discount the utility of his future selves. This argument cannot however ensure that the difference ($d_{0,n+1} - d_{0,n}$) remains constant over time: this would indeed require that Oscar has the same probability of dying at each period, which is highly implausible.

Consider now Parfit's description of Oscar's identity as a sequence of strongly psychologically connected selves. Suppose that the psychological connectedness between two successive selves can be measured by a parameter $0 \leq \beta \leq 1$. We can for instance interpret β as follows: while t_0-Oscar agrees with 100% of the choices he makes today because they are motivated by preferences, values, desires that he considers as being his own), he cannot be sure to agree with more than β % of the choices made by t_1-Oscar, because a fraction (1-β) of the choices made by t_1-Oscar are motivated by preferences, values or desires that t_0-Oscar does not recognise as being his own[5]. Since t_0-Oscar only benefits from €β^n when €1 is given to t_n-Oscar, the psychological distance between t_0-Oscar and t_n-Oscar is $d_{0,n} = (1 - \beta^n)$. We can then deduce the discount factor used by t_0-Oscar:

$$\delta_{0,n} = 1 - [(1 - \beta^{n+1}) - (1 - \beta^n)]$$
$$\delta_{0,n} = 1 - \beta^n (1 - \beta)$$

The discount factor $\delta_{0,n}$ increases with n: t-Oscar therefore discounts his future utilities as if he believed that his future selves would be more patient than him. The psychological distance (as perceived by t_0-Oscar) between selves indeed tends to diminish as n increases. From the perspective of t_0-Oscar, t_{40}-Oscar and t_{41}-Oscar are for instance almost the same person – a person who is quite different from t_0-Oscar. It therefore does not cost anything for t_0-Oscar to impose an important effort on t_{40}-Oscar to the benefice of t_{41}-Oscar.

Reconsider now the decision faced by young-Oscar a bit differently. Oscar is hired for his first job and must decide how much to save for his retirement. He also cares about poverty in the third world and intends to give a part of his salary to a charity. He also knows that he is likely to lose his charitable aspirations while

[5] Frederick (2003) stresses the difficulty of defining an objective measure of psychological connectedness. He for instance measures β by asking to subjects in an experiment to "rate how similar you expect to be in the future compared to how you are now, and how similar you were in the past compared to how you are now. By similar, I mean characteristics such as personality, temperament, likes and dislikes, beliefs, values, ambitions, goals, ideals, etc." on a scale from 0 (completely different) to 100 (exactly the same).

getting older and wealthier. He therefore consciously decides to save a smaller proportion of his current income to be able to give more to the charity today, and imposes on his future selves greater savings efforts, since he cares less about his further selves with whom he does not identify. Within this context, the choice of young-Oscar is not irrational any more: the discount factor he applies to weight his future selves is indeed not a discount factor with respect to time, but a discount factor with respect to psychological connectedness. This implies in particular that the claim that retired-Oscar will regret young-Oscar's choices is not a sufficient reason to nudge young-Oscar today, since it does not mean that young-Oscar's choice was irrational.

LP claims that we should prevent people from being irrational (this is precisely the objective of means paternalism): since it is possible to rationalise time inconsistent behaviours by considering that what matters for *t*-Oscar is not the satisfaction of some stable true preferences but his degree of psychological connectedness with his future selves, we cannot defend nudges on the basis that they help young-Oscar to make better choices for *himself*. Nudging young-Oscar is indeed likely to cause some harm to young-Oscar and to benefit retired-Oscar: the well-being of young-Oscar will however be increased if and only if he has sufficiently strong psychological connections with retired-Oscar.

4. Justifying paternalism

LP and nudges are often justified by claiming that we should protect individuals from their own mistakes while respecting their subjectivity and freedom of choice. The justification of nudges in intertemporal choices (on the basis of personal regret) however requires accepting that (1) people are defined by stable true preferences, (2) they have a correct assessment of their past choices, and (3) they should use a constant discount factor when comparing future utilities. I suggested that conditions (1) and (2) are descriptively inaccurate, while the normative claim of condition (3) – when we assume that a person is unified by the existence of her stable true preferences – implies that the only rationalisable time preferences would be temporal neutrality. I showed that an alternative account of individual identity such as Parfit's idea of psychological connectedness does not necessarily

support the same paternalistic conclusions: it is indeed not necessarily irrational to care less about one's further future if we know that one's identity is likely to evolve over time.

Note that justifying imprudent behaviours (such as saving a small proportion of one's income) on the basis that they are not irrational does not imply that paternalism is not justifiable. Suppose for instance that a reason why young-Oscar prefers to consume more today is that he decides to start smoking. Although this behaviour can be rationalised because young-Oscar is not rationally required to care about retired-Oscar's health, we can argue that young-Oscar directly causes some harm to retired-Oscar (imprudent behaviours can therefore be *morally wrong* (Parfit 1984, p.318)). If intertemporal choices can be seen as choices involving several individuals, we can appeal to Mill's Harm Principle to justify a paternalistic intervention on young-Oscar. A new difficulty then arises, viz. which self of Oscar should be privileged. The issue here is that modelling Oscar's choice as a game between multiple selves with their own preferences (that are nonetheless similar for two relatively close selves) does not provide a concept of welfare that could be applied to the individual as an enduring agent. In a multiple selves model, the determination of a normative criterion on welfarist grounds is therefore not straightforward, since we do not have at our disposal such a notion of enduring agent, and we have no reason *a priori* to privilege a self over another.

It should however be noticed that, unlike within Parfit's analysis of identity, changes in preferences, values, or desires can also be *initiated* by the agent, rather than merely *experienced*. Korsgaard (1989) for instance argues that persons are unified by the continuity of agency of their successive temporal selves, each of them being an active agent, contributing to the shaping of their own identity: what matters from this perspective is not psychological connectedness, but "the view of myself as an agent, as one who chooses and lives a particular life" (p. 23). This argument relies on the Kantian position that we may view ourselves not only as objects of theoretical understanding (the passive *loci* of our experiences) but also as agents, as "the thinkers of our thoughts, and the originators of our actions" (Korsgaard 1989, p.18). What matters for Oscar is therefore the life (in the sense

of long-term commitments) he has chosen, as an autonomous agent, and not the experience of individuals with whom he is psychologically connected.

In this perspective, what behavioural economics tells us is not that real individuals are poor decision makers (who should be helped by the choice architect), but that they lack of autonomy, since they are likely to ruin their long-term plans due to irrelevant framing effects. I consider here a *reasons-responsiveness* account of autonomy, according to which "an agent does not really govern herself unless her motives, or the mental processes that produce them, are responsive to a sufficiently wide range of reasons for and against behaving as she does" (Buss, 2014). The normative issue faced by boundedly rational individuals is not that they are not able to satisfy some hypothetical true preferences, but that their choices can be shaped by reasons they are not aware of (such as framing effects, a present bias… see Hausman and Welch (2010) for a similar argument). Acknowledging the right for the individual to govern herself therefore implies that it is not the satisfaction of one's true preferences that matters, but instead the possibility to choose one's own preferences and identity. Possible measures in this direction would typically consist in educating the individuals by warning them of the existence of framing effects, or more generally of the diverse socio-psychological biases that are likely to affect their choices. A complementary set of measures would then consist in providing to the individuals the means to make and enforce their commitments. So as to offer a philosophically and psychologically coherent solution to cases like Oscar's savings choice, normative economists should probably leave the welfarist background of libertarian paternalism, and focus instead on more deontological criteria.

References

Andersen, S. and Ross, L. (1984). Self-Knowledge and Social Inference: I. The Impact of Cognitive/Affective and Behavioral Data. *Journal of Personality and Social Psychology*, 46:280–293.

Bernheim, B.D. and Rangel, A. (2009). Beyond Revealed Preferences: Choice-Theoretic Foundations for Behavioral Welfare Economics. *The Quarterly Journal of Economics*, 51-104.

Bovens, L. (2009). The Ethics of Nudge. In Grüne-Yanoff, T. and Hansson, S. (Eds.), *Preference Change: Approaches from Philosophy, Economics and Psychology*, pp. 207-219. Berlin and New York: Springer.

Broome, J. (1991). *Weighing Goods: Equality, Uncertainty, and Time*. Oxford: Basil Blackwell

Buss, S. (2014). Personal Autonomy. In Zalta, E. (Ed.), *The Stanford Encyclopedia of Philosophy*. Winter 2014 edition.

Conly, S. (2013). *Against Autonomy. Justifying Coercive Paternalism*. Cambridge: Cambridge University Press.

Frederick, S., Loewenstein, G. and O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature,* 40(2), 351–401.

Frederick, S. (2003). Time Preference and Personal Identity, in G. Loewenstein, D. Read, and R. Baumeister (Eds), *Time and Decision: Psychological Perspectives in Intertemporal Choice*. New York: Russel Sage.

Hausman, D. and Welch, B. (2010). To Nudge or Not to Nudge?. *The Journal of Political Philosophy*, 18(1), 123-136.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: an Analysis of Decision under Risk. *Econometrica*, 47(2), 263-292.

Kahneman, D. and Tversky, A. (Eds) (2000). *Choice, Value, and Frames*. Cambridge: Cambridge University Press.

Korsgaard, C.M. (1989). Personal Identity and the Unity of Agency: A Kantian Response to Parfit. *Philosophy and Public Affairs*, 18(2), 101-132.

Liberman, N., Trope, Y., & Stephan, E. (2007). Psychological distance. *Social psychology: Handbook of basic principles*, *2*, 353-383.

Mill, J.S. (1859). *On Liberty*. London: Longman, Roberts & Green Co.

O'Donoghue, Ted, and Matthew Rabin. (1999). Doing It Now or Later. *American Economic Review*, 89(1), 103-124.

Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Samuelson, S. (1937). A Note on the Measurement of Utility. *Review of Economic Studies*, 4, 155-161.

Samuelson, W. and Zeckhauser, R. (1988). Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty*, 1, 7-59.

Sunstein, C. (1998). Selective Fatalism. *The Journal of Legal Studies*, 27(S2), 799-823.

Sunstein, C. (2014). *Why Nudge? The Politics of Libertarian Paternalism*. Yale University Press.

Sunstein, C. and Thaler, R. (2003). Libertarian Paternalism is not an oxymoron. *Univ Chic Law Rev*, 70, 1159-1202.

Thaler, R. and Benartzi, S. (2004). Save More Tomorrow: Using Behavioral Economics to Increase Employee Savings. *Journal of Political Economy,* 110, S164-87.

Thaler, R. and Sunstein, C. (2003). "Libertarian Paternalism". *AEA Papers and Proceedings*, 93(2):175–179.

Thaler, R. and Sunstein, C. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, New Haven.

Wilson, T. and Gilbert, D. (2003). Affective Forecasting. In Zanna, M. (Eds.), *Advances in Experimental Social Psychology*, volume 35, pages 345–411. San Diego, CA: Elsevier.