THE JOURNAL OF PHILOSOPHICAL ECONOMICS: REFLECTIONS ON ECONOMIC AND SOCIAL ISSUES

Volume IX Issue 1 Autumn 2015

ISSN 1843-2298

Economics of paternalism: the hidden costs of self-commanding strategies

Christophe Salvat

DIU LIBERTAS®
INTERNATIONAL UNIVERSITY

ROSETTI
INTERNATIONAL

Economics of paternalism: the hidden costs of self-commanding strategies

Christophe Salvat

Abstract: The paper proposes an economic assessment of paternalism by comparing different alternative responses to dynamically inconsistent behaviors consecutive to hyperbolic discounting. Two main types of action are possible, self-commanding strategies and paternalism. The first category includes personal rules and pre-commitment. The second can be subcategorized between coercive and non-coercive forms of paternalism, which are respectively associated (although it is debatable) with legal paternalism and with 'nudges'. Despite being self-inflicted, self-commanding strategies are actually not cost free and can result in a dramatic cutback of people's freedom of choice. Likewise, legal paternalism can, on occasion, be less harmful than personal rules or pre-commitment; similarly, nudges can be more invasive and less effective than their proponents want us to believe. The aim of this paper is not to propose any standardized form of response to irrational behavior (whatever that may mean) but to argue, on the contrary, that every case should be individually appraised. Individual situations can be remedied by self-commanding strategies or by paternalistic policies, either in isolation or in combination.

Keywords: libertarian paternalism, personal rules, nudges, self-confidence

Received: 31 July 2015

Salvat, Christophe (2015), 'Economics of paternalism: The hidden costs of self-commanding strategies', *The Journal of Philosophical Economics. Reflections on Economic and Social Issues*, IX: 1, 102 - 124

Introduction

The issue of paternalism has become popular amongst social thinkers over the last decade, and more specifically amongst economists. The success of Sunstein and Thaler's 'Nudge' theory is certainly the best illustration of this revival. Nudge first capitalized on three decades of behavioral studies demonstrating a lack of rationality in individuals and the inability of economic theory to account for this. Economists have found that libertarian paternalism provides them with a way to increase individual rationality whilst preserving consumer sovereignty and thus, more importantly, they found a way to reassert the validity of their theoretical standard of rationality. The philosophical foundations of this theory have often been challenged; however, its economic basis has not yet been thoroughly examined. Whilst Sunstein and Thaler have little difficulty in demonstrating that nudges can have positive effects, they have not yet managed to successfully show that nudges are the most efficient form of paternalism, or that paternalism is even the best response to people's imprudence. The present article addresses this issue and claims that nudges are only one alternative amongst many, and that a careful cost-benefit analysis is needed before generalizing this kind of action.

This paper is structured around six sections. The first section narrows the discussion to very specific types of imprudent actions, those generated by variable time discounting rates. This allows us to avoid questioning the rationality of individuals' preferences, and to restrain the possible field of paternalistic actions to actions that individuals themselves would prefer not to undertake. It does not necessarily mean that paternalist or non-paternalistic interventions are not justified in other cases or that they are always justified in that type of case. Limiting the scope of the discussion to a commonly accepted type of irrational behavior is solely a way of focusing on the comparative advantages of alternative responses to irrationality (whatever irrationality actually means). The second section presents the first possible response to imprudence, i.e. self-commanding strategies, which either refer to personal rules or to pre-commitment. The third section discusses the possibility of paternalistic actions, which can either refer to allegedly coercive actions (legal paternalism) or allegedly non coercive actions (nudges). The fourth section compares the relative cost (in terms of freedom of choice) of these possible solutions. The last section concludes this study by

rejecting pro and anti-paternalism supporters back-to-back and suggests a differentiated approach based on pragmatism.

Opportunities of paternalism

In this section I consider the conditions necessary to make paternalism justifiable. There are many approaches to this question. One consists of stating that paternalism is never justifiable, however large the benefits and however low the cost might be for the person. This is a position that I shall not consider here since my aim is to assess whether, and which, cases of paternalism can be economically justifiable.
I do not intend this study to be a defense of paternalism. I would rather prefer to condemn paternalistic actions on their relative inefficiency and/or cost than for pure ideological reasons. A second approach consists of saying that paternalism is justifiable whenever it is beneficial to the people interfered with. Although partisans of paternalism rarely explicitly endorse this view, I believe it is the most common amongst them. One can argue the belief that benefits should always be related to their potential cost, whether immediate or long term. And the main cost of paternalism is, as we know, the reduction of personal autonomy. There might be other negative consequences, but I shall only consider this one in the present study.

There are two ways of assessing the benefits of paternalism. The first is to objectively claim that individuals would be better off with an external intervention than without and the second is to trust individuals in their own appreciation of what is good for them. Economists have a strong preference for the second one, and it is often due to this point of view that paternalism is discarded as being totalitarian. Both approaches have their own merits and weaknesses. Individual preferences are not always 'rationally acceptable' even if they respect the conditions of completeness and transitivity set by economists, and we do not all share a same 'objective' view of individual well-being. Considering both positions, I believe the preference approach to be more reasonable than the objective approach. I shall therefore adopt here the standard economic view, according to which individuals are rational in so far as they satisfy their own preferences (even if I am aware that *some* preferences are intrinsically irrational). A second reason motivates this choice: partisans of an objective approach to well-being are usually more sympathetic to the idea of

paternalism than the defenders of the subjective approach. If I were to show that paternalistic actions can be more efficient and/or less costly than non-paternalistic actions, my point would be even stronger had I demonstrated it from a traditionally anti-paternalistic point of view.

Let us now consider cases which could *potentially* qualify for paternalistic actions. My main concern here is to identify potential cases of inconsistent behavior. I propose to focus here only on one kind of inconsistency, preference inconsistencies over time or Dynamically Inconsistent Behavior (DIBs).

Dynamically Inconsistent Behaviors are extremely common in the population and are often caused by hyperbolic discounting. DIB's have been associated with a wide range of expected utility anomalies from procrastination (Akerlof, 1991; O'Donoghue and Rabin, 2001) to credit-card debt (Laibson, 1997) or addiction (Gul and Pesendorfer, 2007; Herrnstein and Prelec, 1992).Discounted-Utility models are classically exponential, using a reducing factor of $1 / (1 + k)^t$ where k is the constant discount rate and t is the length of the delay. Given a constant discount rate, the discounting becomes exponential. The value of future rewards directly (and exclusively) depends on the length of time that individuals have to wait for them. Unless the discounted value of rewards expected in t_n exceed the non-discounted value of present rewards, individuals are considered perfectly rational in preferring present to future benefits. Preferring the present to the future, therefore, is a matter of personal taste rather than an inconsistent inter-temporal choice. Exponential discounting models *a priori* exclude any possibility of paternalism as they assume the preference for the present is voluntarily chosen. Any attempts to prevent a person's future harm is not only illegitimate (as it violates individual freedom) but also irrational (as it does not maximize utility). This is the position adopted by Gary Becker in his works on addiction for instance (Becker and Murphy, 1988). On the contrary, discounting rates are varying in cases of hyperbolic discounting. The expression 'hyperbolic discounting' is actually slightly misleading since time discounting functions are not strictly hyperbolic but quasi-hyperbolic (Laibson, 1997). Their first axiomatic analysis was presented by Robert Strotz in 1955-56 (Strotz, 1955).

$U^t(u_t, u_{t+1}, \ldots, u_t) = \delta^t u_t + \beta\Sigma^{\tau}_{\tau-t+1}\delta^{\tau}u_{\tau}$

Where $0 < \delta, \beta \leq 1$

If $\beta = 1$, then exponential discounting preferences

If $\beta < 1$, then (quasi) hyperbolic discounting preferences

In hyperbolic discounting the bias is introduced by the sense of proximity of the delay (and not just its duration). The closer it gets the higher the time preferences go. Hyperbolic discounting functions use a multiplying factor $\beta < 1$, reducing discounting rates as the discounted event moves further away in time. Events located in a near future are discounted at a higher discount rate than events located in a distnt future. Future rewards are reduced by a factor of $1/(1 + kt)^{\beta/\alpha}$ where α and β are greater than zero. The longer is the delay, the lower is the discount rate. Alternatively, the closer is the expected reward, the more discounted is the expected event. What makes this bias so specific (and dynamically inconsistent) is that a person's rate of discounting increases as the deadline approaches. As a matter of fact, discount rates change with the passage of time, as individuals reevaluate their plans accordingly, abandoning the plans they made in the past and making new plans for the future. Plans that had been rationally elaborated in the past with a long term discount rate cease to be optimal once the deadline approaches and a (higher) short-term discount rate is applied. Periodically reevaluating one's plans logically leads to long-term utility washouts. The latest plans are, in that respect, irrational.

More interestingly for our purpose, the closer the deadline gets, the less desirable the next option becomes, to the point where the last one is reluctantly chosen by the agent. Take the following example presented by O'Donogue and Rabin. 'Suppose you usually go to the movies on Saturdays, and the schedule at the local cinema consists of a mediocre movie this week, a good movie next week, a great movie in two weeks, and (best of all) a Johnny Depp movie in three weeks. Now suppose you must complete a report for work within four weeks, and to do so you must skip the movie on one of the next four Saturdays. When do you complete the report?' (O'Donoghue and Rabin, 1999: 109). The best plan, all things considered, is naturally to complete your report the first week and to only miss the mediocre film. But, given your time-biased preferences, you

prefer going to the cinema the first week and commit to write the report the following week. Being time-biased (and being completely naïve about it, I shall come back to that later), you choose to indulge on the first Saturday off (even if it is to see a mediocre movie) and commit yourself to write the report in the following week (in order to be still able to see the great movie and the Johnny Depp movie, which has your preference). The following week, however, your discount rate having increased, you logically reconsider your previous choice, and go to the cinema again. The same thing happens in week 3 obliging you to complete the report on the last Saturday and to miss Johnny Depp's latest movie. Missing Johnny Depp's latest movie to write a report is clearly not a voluntary choice.

To fully appreciate the effects of time inconsistency, one must therefore take in consideration a further fact: sophistication. Whereas some people are fully aware of their time inconsistency (due to their past experiences for instance), others seem genuinely taken aback each time it happens. Strotz initially distinguished two types of individuals, the 'spendthrift' individuals who do not recognize their dynamic inconsistency and the 'thrift' individuals who do (Strotz, 1955). They are now more commonly referred as 'naïf' and 'sophisticate'. Let us keep the latter terminology. The distinction is an important one. In terms of welfare, naïfs usually suffer from bigger utility losses than sophisticates. While sophisticates (even incorrectly) anticipate their preference reversal, and (eventually) take action using self-control or external commitment device (see next section), naïfs repeatedly choose dominated options, endlessly incrementing their utility losses. Even a small hyperbolic bias (β close to 1) can have dramatic effects on the naïfs' long-term welfare, whilst they are theoretically harmless for sophisticates. Paternalism, consequently, should primarily aim at stimulating sophistication. I strongly emphasize the fact that, contrary to common beliefs on imprudence or irrationality, the absolute magnitude of the person's discount rate has no bearing *per se* on her utility loss. She is not imprudent because she discounts too much her future rewards but because she discounts them more as they get closer. In fact, over long periods, hyperbolic discounting functions generally tend to discount future rewards at a lower rate than exponential ones.

Differentiating naïfs from sophisticates can be useful to assess the legitimacy of paternalistic actions [1]. It is indeed easier to identify a person's real preferences when she is sophisticated rather than naïve. Even if they fail to satisfy their initial

preferences, the fact that sophisticates expected a future preference reversal and took step against it is a testimony of their real preferences. This means that their action is involuntary, and henceforth eligible to paternalistic policy. Such is not the case of naïfs, who systematically deny their initial and utility-maximizing choice when they fail to implement it. When it comes to paternalism, the result is puzzling. On one hand, unlike sophisticates, naïfs are badly hurt by small hyperbolic bias. On the other hand, paternalistic actions can be better justified in cases where individuals are sophisticate than naïve. That ironically makes paternalism most legitimate when unnecessary.

Self-commanding strategies

In the following, I shall discuss the potential responses to Dynamically Inconsistent Behaviors (DIBs): paternalistic and non-paternalistic policies. This will then enable me to respectively compare their relative cost. In the present section, I start with self-imposed policies only accessible to sophisticated agents.

One of the most common liberal beliefs is that it is always better to let individuals choose for themselves. Let us suppose for now that individuals are sophisticated. Despite being spontaneously inclined to overvalue earlier rewards relative to later ones, sophisticated agents can show a remarkable economic rationality. A first way to deal with DIBs is to adapt. This is the strategy of consistent planning initially described by Strotz. Consistent planning is a bargaining game, played between the present and future selves of a single agent. Plans that her present self knows will be overturned by her later self are definitively discarded so that a subgame equilibrium (known as the Strotz-Pollak equilibrium) is to be found amongst the remaining feasible options (Pollak, 1968; Peleg and Yaari, 1973; Goldman, 1980). Despite being freely chosen, this solution remains highly unsatisfactory since it does not really address the underlying issue.

Two other options can be considered, personal rules and pre-commitments. They constitute the main self-commanding strategies used to counteract DIBs. While personal rules are purely self-commanded, pre-commitments need the mediation of external individuals. Let us first consider intermediated self-commanding strategies.

Intermediated self-commanding strategies usually take the form of a contract. Ulysses demanding his crew men to bind him to the ship's mast when approaching the sirens is a classic example (Elster, 1979). There are many other ways to trump temptations, many of which do not involve physical constraints. Schelling listed a number of them, among which relinquishing authority to somebody else, committing or contracting, disabling or removing yourself, removing mischievous resources, submitting to surveillance, incarcerating yourself, arranging rewards or penalties, rescheduling your life and setting yourself the kinds of rules that are enforceable (Schelling, 1984: 6-7; Schelling, 1992). But the most common cases either rely on a strong moral binding, like a promise, or on a short-term game plan like a pre-agreed utility cost (Gul and Pesendorfer, 2001; Gul and Pesendorfer, 2004). Deadlines (and the moral or financial sanctions that go with them) are notoriously useful for coping with procrastination. Similarly, gyms do not thrive in spite of, but because of, their rigid and expensive membership terms. Note, however, that third parties who are implicated within these self-imposed rules as part of the enforcement mechanism are not necessarily mandated for this task and are sometimes not even aware of their mediation role. Social blame and public reprobation can be a deterrent as effective as any financial penalty. This is the strategy adopted by members of Alcoholics' Anonymous for instance.

Pre-commitments are usually highly effective but they present at least two important drawbacks: they require foresight and they lack flexibility. Like any self-commanding strategies, pre-commitments are only open to sophisticated individuals. This, in itself, considerably reduces their scope. But, in addition to that, they are necessarily restricted to expected events. Ulysses escaped his fate because he precisely knew where the sirens would be. Had he not known, he would have had to stay bound to his mast throughout the whole duration of the Odyssey. Knowing one is time biased is one thing, but knowing when one will need control is another. Let us assume that one knows when control will be needed, and that a contract has been accordingly agreed, there is still the possibility for the contract to be overly or insufficiently strict. If the contract is insufficiently strict, it will be ineffective. If the contract is excessively strict, then it can be unnecessarily harmful if not dangerous. Consider, for instance, the case of removing 'mischievous' resources, discussed by Schelling (Schelling, 1984). A woman expecting her first child has decided to fully and consciously experience the birth of her baby: she wants to deliver without any anesthetic. The physician proposes to have an anesthetic ready for her to use in case she needs it. She knows, and the experience of

others show (Christensen-Szalanski, 1984), that, with the prospect of the pain drawing nearer, she will reconsider her choice, and ask for pain relief, a choice that she thinks she will regret later when the baby is born. She therefore refuses the physician's proposition and demands all anesthetic to be removed from the room when she gives birth. If the physician keeps the anesthetic close at hand, it is more likely that she might use it. But if the physician does not offer her the possibility of using it at all, then she might experience an unbearable pain, and never want to give birth to another child ever again. Pre-commitments are instrumental in our well-being but it is irrational to use them at any cost.

Non-mediated self-commanding strategies, commonly known as personal rules, offer more flexibility. Personal rules can be implemented without any external help and give (at least at first sight) sophisticated agents more latitude to deal with their own inconsistencies. Individuals typically implement personal rules to stop smoking, avoid procrastination, go on a diet or work harder, all problems raised by DIBs. Unlike pre-commitments, personal rules are often kept quiet, either because individuals are ashamed of their problems or because they are ashamed that they are not able to solve them and could appear weak. Note that when personal rules are made public, it is generally to use public reprobation or personal shame as an additional incentive to obey one's rule. Personal rules can thus easily turn into pre-commitments. When they are kept secret, and nobody but the individual can see that they are being broken, enforcing them becomes a real challenge. Diets are suspended on bank holidays, then on birthdays, then on week-ends, and then every time one goes to the restaurant or one has a pizza. Similarly, cigarettes are first tolerated in stressful times, and end up being smoked on all occasions. For having witnessed it around us, and for having personally experienced it, we all know that personal rules, when they are kept personal, do not generally work.

To be effective, personal rules require willpower. Willpower is, by definition, the ability to resist temptation (Ainslie, 1992: 142-143). This is, as we all know, rather unequally distributed among the population. Some people are lucky enough to show a startling level of willpower (or so it seems), when others are victims of their incapacity to resist natural impulses. And there are many ways to improve one's willpower (Holton, 2009; Bénabou and Tirole, 2002; Bénabou and Tirole, 2004). But appearances can be also deceptive: seemingly strong-willed individuals can also turn out to be weak

but severely self-constrained. Willpower is commonly believed to be a natural feature. It may well be to some extent. But it is mostly acquired through experience, education and good faith. The economists Roland Bénabou and Jean Tirole showed that willpower essentially works as a *reputational capital*, growing each time a personal rule is successfully tested and decreasing each time it fails the test (Bénabou and Tirole, 2004). Individuals are not obliged to test their willpower, in which case they indulge their current impulses without losing confidence in their willpower. For Bénabou and Tirole, individuals who already have a certain reputational capital to preserve have therefore an additional incentive to comply with their personal rules, making strong-willed individuals even more likely to stick to the rules. Conversely, weak-willed individuals who have difficulty in resisting temptation are getting weaker and weaker as their rules lapse. Two factors explain this vicious circle. Firstly, the smaller an individual's reputational capital, the smaller their incentive to protect it. Secondly, once a rule has been broken, it significantly loses its moral power. Lapsing once to a rule creates what Ainslie calls a 'precedent', and makes any further lapse easier to happen [2]. Partners, for instance, usually take a vow not to cheat on each other. The vow is explicit for married couples, but is usually implicit for unmarried couples. When opportunities to cheat turn up (as they often do), they are usually dismissed thanks to the moral cost (the shame) that would inevitably result from breaking this personal engagement. But once one partner has cheated then the harm is done and there is no further disutility in cheating a second time, and then a third etc. The (moral) cost of breaking the rule is mostly entirely born with the first lapse. Cheating again only marginally adds to the guilt felt the first time. According to this logic, strong-willed individuals naturally tend to get stronger when weak-willed ones naturally tend to get weaker. This is, however, not necessarily the case. Naturally weak-willed individuals can also turn out to be the most rigid rule followers.

This claim can seem surprising at first. Psychologist George Ainslie brilliantly demonstrated that precisely because individuals were weak-willed or believed themselves to be so, they were prone to take drastic steps to avoid future failures. If they want to resolve their problems alone, weak-willed individuals must be uncompromising with themselves. They usually start by denying themselves any excuse to lapse to the rule. People who systematically follow their rules and seem to be strong willed may in fact be under confident and therefore uncompromising in their behavior. They are simply overly harsh with themselves, punishing themselves when they lapse,

or even worse, externalizing the punishment. This is what Schelling calls side-betting (Schelling, 1960). Side-betting differs from commitment since commitment requires the explicit consent from a third part, when side-betting does not. One way of externalizing punishment is typically to submit oneself to public reprobation or shame. This is the solution adopted by members of Alcoholic Anonymous. Some determined (yet weak-willed) individuals are more imaginative. Take for example the case of the drug addicted physician quoted by Schelling: 'In a cocaine addiction center in Denver, patients are offered an opportunity to submit to extortion. They may write a self-incriminating letter, preferably a letter confessing their drug addiction, deposit the letter with the clinic, and submit to a randomized schedule of laboratory tests. If the laboratory finds evidence of cocaine use, the clinic sends the letter to the addressee. An example is a physician, who addresses a letter to the State Board of Medical Examiners confessing that he has administered cocaine to himself in violation of the laws of Colorado and requests that his license to practice be revoked. Faced with the prospect of losing his career, livelihood, and social standing, the physician has a powerful incentive to stay clean.' (Schelling, 1992: 167). This case is exemplary of the extreme means that some people are willing to use to overcome their weakness. Ainslie was the first to demonstrate that a poor perception of one's willpower, or too high an image of other people's willpower, often leads weak individuals to use excessively harsh means. Men and women dread flexible rules as they dread their own weakness. This explains why some people prefer to stop socializing altogether when they go on a diet, or why others turn into workaholics to avoid procrastination. Compulsion succeeds to personal rules, and prudence gives way to miserly behavior and bigotry.

However, self-commanding strategies are not always that costly. They are still widely used with success or near success in everyday life. Everyone lapses now and then, to one's personal rules, but overall we rather successfully comply with them. And in the vast majority of cases, these lapses are not significant enough to call for extra sanctions. Did I eat a second *pain au chocolat* this morning when I only intended to eat one? Well, I'll try to be more cautious next time. That's it. Did I oversleep this morning? Well, I'll put an extra alarm clock on my phone for tomorrow. There is no need to pre-commit, or to venture into complex and harsh side-betting. Those strategies should be reserved for individuals who are absolutely unable to follow a rule, and to cases in which lapses have serious consequences that cannot be forgiven.

Possibility of paternalism

In the previous section I presented non-paternalistic solutions to DIBs. Although many, if not most, cases of spontaneous responses to hyperbolic discounting can be resolved without external interventions (Holton, 2009), others, however, may require such action to be taken. The use of external intervention is normally related to cases where individuals are naïfs and therefore unaware of their dynamic inconsistency, or when the self-imposed solutions of sophisticated individuals are too costly in terms of autonomy. In these situations, paternalism might then be considered. Paternalistic actions can be characterized as unsolicited actions undertaken by a benevolent agent and motivated by someone else's irrational or imprudent behavior. They specifically target the decision-making process by checking or by boosting the other individual's motivation. This is either achieved by emotional or psychological tricks or, alternatively by means of authority. By unsolicited actions, I mean all actions that have not been explicitly called for. That naturally includes, but is not limited to, coercive actions. Unsolicited actions also include unsolicited advices or unsolicited help. Unsolicited actions are by definition not pre-consented, but they can be endorsed *a posteriori*.

Paternalism has been traditionally discarded as a harmful hindrance to individual freedom (Dworkin, 1983: 142-143; Feinberg, 1986; Marneffe, 2006). These authors were essentially concerned with the loss of individual freedom potentially resulting from paternalistic actions. This is indeed a major moral consideration, although actually not the only one (Salvat, 2014). Recently, Sunstein and Thaler argued that paternalism can be as innocuous as a nudge, hence opposing non-coercive paternalistic 'nudges' with traditional legal – and coercive – paternalism. Their argument is, however, somewhat misleading. It could be argued first that making a coercive action less visible or painful does not mean that the action is any less coercive. Manipulation can also be a form of coercion (Bovens, 2009). Paternalistic laws, secondly, can claim to be freely chosen in democratic countries. Paternalism can therefore be a collective choice, as Rawls himself admitted (Rawls, 1971). I therefore propose to judge paternalistic actions not as what they seem, but on what they actually cost to individuals in terms of freedom of choice. In the previous section, I showed that self-inflicted rules or pre-commitment could be costly; I want to see here how paternalistic actions compare to this.

How are we to assess the relative cost of paternalistic policies? As I stated at the beginning of this study, I am only interested here with paternalism's impact on individual freedom. So by cost one should understand infringement of freedom of choice. It is also clear enough by now that paternalistic policies are assessed relatively to non-paternalistic policies, which also have a cost (personal rules, pre-commitment, side-betting etc). Last, but not least, I consider that the severity (and therefore the cost) of paternalistic and non-paternalistic policies not only depends on the agent's actual willpower but also on the extent of her subjective (and often mistaken) appreciation of her own willpower. It is therefore important to distinguish willpower and self-confidence. I shall refer to willpower as the probability to comply with the rule previously adopted or the decision previously taken and use self-confidence to refer to the perception an agent has of her own willpower. Hence an agent is said to be over-confident when, for instance, she believes her probability of sticking to her rule or her decision is ¾ when it is actually ¼. Conversely an agent is said to be under-confident when she believes the probability that she will stick to her rule is ¼ when it actually is ¾. An agent who has an exact appreciation of her willpower is fully sophisticated.

I can now offer a (rough) typology of paternalistic actions that are made possible in the case of DIBs:

- Lack of foresight or bad luck (P_1)
- Lack of sophistication (P_2-P_3)
    - agents fail to comply with their personal rules (P_2)
    - agents commit themselves to unreasonably harsh or demanding rules (P_3)
- Absence of sophistication (P_4)

Let me develop them in order.
P_1 refers to cases in which individuals, although sophisticated, are unable to resort to self-commanding strategies. Sophisticated agents, for instance, know that they are time inconsistent, but they fail to correctly identify situations in which their time bias will be an issue. This is a problem in particular for pre-commitments, which require complete and accurate anticipation. Consider the case of Ulysses. Ulysses knows where the sirens will be and is bound accordingly. He is free for the rest of the journey. Consider, next, the case of Moby Dick's Captain Ahab, initially quoted by Schelling

(Schelling, 1984).In the film (this particular episode is not in the book), Ahab seriously injured his leg in the water. Back on the boat, Ahab is held down by his crewmen so that the blacksmith can cauterize the stump with a hot iron. Ahab refuses to be burned. Ahab is in no way different from Ulysses except in that that he did not expect his predicament. Had Ahab anticipated his accident, he would have instructed his crewmen in one way or the other beforehand. Personal rules are a more appropriate response to this kind of situation but, as illustrated by Ulysses' tale, they are not applicable to all cases.

P_2 and P_3 refer to cases in which personal rules have been implemented but are either insufficiently effective or too costly. Let us first consider the case of insufficiently effective personal rules (P_2). Failures to comply with personal rules are usually believed to be the result of weak willpower. There is actually more to that. What is really decisive is how well you appreciate the level of your willpower rather than the level of your willpower itself. Let me give you an example. Suppose there is a customer with a willpower level of $^1/_4$. She goes each day to the cafeteria, and each day she has the choice between fruit salad and cheesecake. Her preference, before going to the cafeteria, is always to take fruit salad. But having a level of willpower of $^1/_4$ the probability of her taking the fruit salad is only one of every four occasions. Being sophisticated to some extent, she does not know her real chances but she knows that she has some chance of taking the cheesecake in the cafeteria. Imagine first that she overestimates her willpower. Suppose that she estimates her chances of choosing the fruit salad at $^3/_4$. Because she believes her chances to be already quite high, she gives herself a relatively soft but rather ineffective rule. Paternalistic solutions are then possible (P_2). They are not necessarily justified (see next section) but there is also an opportunity here. Alternatively, the customer underestimates her willpower, and sets a personal rule that, although very effective, can be unduly costly (as illustrated by Schelling's cocaine addict physician). In those cases, paternalism can potentially be offered as a less coercive solution (P_3).

Table 1: Possibility of paternalism

Self-confidence	$^3/_4$	$^1/_4$	$^1/_8$
Willpower	$^1/_4$	$^1/_4$	$^1/_4$
Personal rule	Insufficient	Adequate	Excessive
Paternalism (possibility of)	P_2	None	P_3

Finally, P_4 refers to cases where self-commanding strategies are simply not an option due to individuals' naivety. This is a theoretically straightforward case that supporters of paternalism are eager to use. In practice, however, P_4 raises a number of ideological issues. Apart from some particular cases, it is indeed debatable whether people can be entirely unaware of being time biased. Mentally deficient adults and young children undoubtedly belong to this category and need external guidance. For the rest of the population, I am personally inclined to believe that, although a lot of people are excessively over-confident, very few are completely naïve. The opposition introduced by Strotz between sophistication and naivety is fundamental but one would be misled to oppose pure sophisticated and pure naïve individuals. The vast majority of people actually belong to an intermediate category. P_4 is nonetheless possible.

Price of paternalism

Let me now assess the cost of paternalism to each of these cases separately. P_1 and P_4 are particular in the sense that they do not allow comparisons between paternalistic and non-paternalistic responses to hyperbolic discounting. Since it is my aim in this article to assess the cost of paternalism relative to self-commanding strategies, I shall exclusively focus on P_2 and P_3.

Table 1 showed that paternalism is either made possible by excessive or insufficient self-confidence. It is now our responsibility to find out when these paternalistic actions can be actually justified. As I explained above, I believe paternalism is neither intrinsically legitimate nor illegitimate, but that its legitimacy depends on its relative cost. And by cost, I mean the reduction of freedom of choice involved by its implementation. It has been sometimes argued that paternalism does not necessarily

infringe people's autonomy, autonomy being understood as the expression of the person's real will. It is a defendable point of view but one that has no chance being heard by libertarian thinkers. I therefore propose to assess the relative impact of paternalistic and non-paternalistic solutions to hyperbolic discounting in terms of choice rather than autonomy.

I also argued in the previous section that the efficiency of personal rules depended not only on willpower, as commonly presumed, but more importantly on self-confidence. The higher the gap is between willpower and self-confidence, the harder the intervention needs to be. Individuals generally set themselves a rule whose severity is inversely proportional to their self-confidence. The more confident they are, the more innocuous is the rule they set for themselves. But, at the same time, the weakest is their willpower, the harder the rule needs to be. The difference represents the relative cost of paternalism (Table 2).

Table 2: Cost of Paternalism

Agents	A	B	C	D	E
Self-Confidence	1	$^4/_5$	$^3/_5$	$^2/_5$	$^1/_5$
Willpower	$^2/_5$	$^2/_5$	$^2/_5$	$^2/_5$	$^2/_5$
Personal rule (severity)	Nil	Soft	Medium	Adequate	Too Hard
Intervention (relative cost of)	High	Medium	Low	Nil	Negative

Agent A thinks that she does not need a personal rule because she believes that she is strong enough to avoid any preference reversal. Yet being actually weak, she does need a rule. Because she failed to see that, the rule that might be imposed on her will be resented as extremely costly. B and C are also over-confident but they are sophisticated enough to adopt a personal rule. Because B's rule is too soft relative to her willpower, she might need – if she really wants to achieve her aim – additional self-commanding or paternalistic devices. Note that B and C do not necessarily need paternalism. In many cases, occasionally lapsing rules are of no great consequence. Furthermore, individuals progressively adapt their rules to their actual willpower. But if paternalistic actions are nevertheless undertaken, their relative cost should be

moderate. The more the severity of the paternalistic rule approaches the severity of the personal rule, the less it is resented as infringing one's freedom of choice. When they are equally severe, the relative cost of the paternalistic intervention is nil (D). This is what happens when the agent correctly evaluates her willpower. The paternalistic rule (the law for instance) then perfectly corresponds to the rule she would personally adopt. Despite being externally imposed on her, the law does not generate extra cost for her to follow. In other words, a same rule (say a law forbidding selling extra-large cups of sodas) can be resented as coercive by some people (like A) whilst others (like D) will not suffer from it. One can even go further as some people (like E) might actually see their freedom of choice partly restored by the paternalistic rule. Under-confident agents are, as we saw in a previous section, inclined to set unnecessary harsh personal rules and/or to sanction them with side-betting. Consider Stephanie, who is well aware of the risks generated by drinking too much. She is willing to reduce her consumption and to avoid the temptation of buying extra-large cups (despite them being cheaper). However, Stephanie – like E – does not trust herself with an 'Avoid buying extra-large cups of soda' rule and forces herself to adopt a 'Never buy sodas ever again' rule. When the law forbidding selling extra-large cups is enforced, she can dismiss her personal rule, and enjoy buying sodas again. In very specific cases, seemingly coercive legal rules can foster – rather than infringe – individual freedom of choice. Paternalistic strategies are not necessarily more coercive than their non-paternalistic alternatives.

I have shown so far that infringement of individual freedom is not a general feature of paternalism, and that, contrary to some common belief, paternalism is not necessarily more coercive than self-command and even can, in some cases, be a better alternative. I have shown, in particular, that much actually depends on the agents' level of sophistication and her degree of self-confidence. I would like now to elaborate on the relationship between legal paternalism and nudges, on one hand, and legal paternalism and personal rules, on the other hand. My aim here is to show, based on the results found so far, that it is misleading to oppose these different policies, as they usually work best together.

Table 2 shows that nudges are only efficient in intermediate cases, while legal paternalism is most beneficial in more extreme cases. Take, once more, the example of the cafeteria and suppose that customers have an initial preference for healthy food. Rearranging the dishes will have no effect on A, D or E. There is no point in nudging

A as she does not even have a personal rule to follow. As for D and E, the rule they imposed on themselves is sufficiently harsh (and even excessively harsh in the case of E) for them to prefer the healthy food whatever the dish arrangement is. B and C, on the other hand, can be affected in their choice by the line rearrangement. Because her willpower is only slightly overestimated, C is more likely to be nudged than B. So out of the five cases presented in Table 2, only C (and possibly B) can potentially be nudged [3]. On the other hand, A and E can only benefit from rigid and seemingly coercive rules. In the case of A, preference reversals are unavoidable even in the presence of nudges. Personal rules and self-commitment are also made impossible by the agent's extreme self-confidence. In a case like this, the only possible way to remedy DIBs (if this is justified) is to impose rigid rules. In the case of E, as explained above, laws are not only beneficial to solve time inconsistencies but they also have a negative cost compared to personal rules. A first conclusion is then that, contrary to what Sunstein and Thaler claim, nudges cannot be generalized and they certainly cannot be an alternative to legal paternalism. At best, nudges help slightly overconfident people.

A second important point is that personal rules and external ruling (or nudging) are complementary rather than alternative policies. As explained above, the relative cost of paternalism (in terms of opportunities of choice) functionally depends on the gap existing between the agent's actual and perceived willpower. Ideally, paternalistic laws should be modeled on the personal rules perfectly sophisticated agents – like D – give themselves. The more people deceive themselves the more costly the law becomes. In other words, the coerciveness of the law is the price of self-deception. But laws do not only sanction self-deception, they also contribute to improving people's sophistication. Paternalistic laws, when they are properly set and when all concerned know that they represent the personal rule that they themselves should have, give each agent the possibility to infer from it the extent of her self-deception. If all individuals were fully rational, they would use paternalistic laws to revise their level of self-confidence, until their personal rules and the paternalistic laws corresponded perfectly. Whether they actually are rational or not is another question.

To finish, I want to emphasize the limitations of this study. As I made clear at the beginning, my aim here is not to defend paternalism but to consider – as objectively as possible – the cost of paternalism relative to self-command. I conclude that paternalistic actions are not necessarily more costly than self-commanding ones, and

that legal paternalism can help overconfident agents to establish better personal rules. I have based my conclusions on a single example, supposing a common willpower, and different levels of self-confidence. The degree of willpower chosen $^2/_5$ is voluntarily low to make the case more compelling. Likewise, I implicitly supposed that the case in question was serious enough to possibly prompt legal action. Yet I certainly do not believe that all instances of DIBs call for paternalistic actions (legal or not). To emphasize, my only purpose here was to assess the relative cost of paternalistic actions compared to non-paternalistic ones *when an action is needed*. It is clear that all time inconsistencies (since I specifically focused on those) do not call for an action and that some of them can possibly call for a personal action but not for a paternalistic one. It would certainly be helpful to know when paternalistic actions are socially tolerable before assessing them. Ultimately, I leave this question open for future debate.

Conclusion

I claimed in this paper that the cost of paternalism, generally speaking, should not be assessed in absolute terms (in terms of opportunities of choice, for instance) but *relatively* to the cost of self-commanding strategies. Nudges may not be very constraining but they can be relatively more constraining than personal rules. Conversely legal rules can be coercive but comparatively less so than personal rules. If freedom of choice is to be adopted as moral criterion, paternalist policies should then be tested with comparative benefit-cost analysis. Comprehensive studies would undoubtedly show that legal paternalism is sometimes to be preferred to any other alternatives. There are different types of responses to irrational self-regarding actions, of which self-command, behavioral paternalism, legal paternalism, empathic paternalism (a type of paternalism I unfortunately had no room to investigate here) etc. None are absolutely good or better than the others, including self-commanding strategies which are highly regarded by liberal thinkers, but they all have a role to play. I argued that Sunstein and Thaler's paternalism is useful in specific cases but that it cannot be a substitute for legal paternalism or for personal rules. Nudges can improve the efficiency of personal rules but they are pointless in their absence. Legal rules, on the contrary, can (and ought to) stand in for personal rules when they are inexistent or when they are inadequate. In the great majority of cases, however, legal

and personal rules work together. Legal rules play an important role towards sophisticated people: they set a standard people can easily refer to.

Endotes

[1] A second important distinction consists in telling apart the respective reactions of sophisticates and naïfs before immediate rewards and immediate costs. O'Donoghue and Rabin (1999) show that it is wrong to assume that naïfs are always worse off than sophisticates. If sophisticates are better equipped to fight off procrastination, for instance, they are prone to go too far the other way, preproperation. In a number of situations, such as this one, sophisticates' attempting to outbalance their time inconsistency can result in excessive prudence, or even negative time discounting. Sophistication is generally an advantage when it comes to time inconsistency, but self-medicated policies can have a substantial cost that ought to be included when assessing the relative (dis)utility of paternalism.

[2] One could argue that individuals can 'forget' past lapses and that this would therefore not affect their reputational capital. It is difficult to know exactly to what extent these memory lapses are conscious (bad faith) or unconscious. In any case they are very common, and can significantly counter-balance or exacerbate the self-enforcing nature of personal rules. On one hand, unremembered achievements do not contribute to bolstering reputational capital and ignored failures do not jeopardize it. If 'memory gaps' were equally distributed between successes and failures, the trend observed above would not be disproved. Yet empirical studies have shown that human beings are more likely to forget their failures than their successes. A logical consequence of this is that weak-willed individuals do not get weaker, but that they merely remain weak, whilst strong-willed individuals still get stronger. The gap between weak and strong willed individuals is still deepening but at a slower rate. On the other hand, selective memory can extend to testing one's willpower. The difference with the previous point is slight but important. In the previous case, weak-willed individuals forgot that they failed the test. They are nonetheless still aware of their time inconsistency. In the present case, weak-willed individuals forget even that they put their willpower to the test. This implies that they are now denying their preference reversal. Like the fox in front of the sour grapes, they prefer rationalizing their failure

rather than admitting it or even forgetting it. Their self-confidence remains intact, and they have now nothing to envy in the strong-willed individuals.

[3] It does not, however, mean that opportunities to nudge are relatively scarce. It could well be that 80% of the population belongs to categories B or C.

References

Ainslie, George (1992), *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*,Cambridge, UK: Cambridge University Press.

Akerlof, George A. (1991),'Procrastination and Obedience', *The American Economic Review,* 81 (2), 1-19.

Becker, Gary S. and Kevin M. Murphy (1988),'A Theory of Rational Addiction', *Journal of Political Economy,* 96 (4), 675-700.

Bénabou, R. and J.Tirole (2002),'Self-Confidence and Personal Motivation', *The Quarterly Journal of Economics,* 117 (3), 871-915.

Bénabou, R. and J. Tirole (2004),'Willpower and Personal Rules', *Journal of Political Economy,* 112 (4), 848-886.

Bovens L. (2009),'The ethics of nudge' in Till Grüne-Yanoff and Sven Ove Hansson (eds), *Preference Change. Approaches from Philosophy, Economics and Psychology*, Netherlands : Springer, pp. 207-219.

Christensen-Szalanski, JJ. (1984),'Discount Functions and the Measurement of Patients Values', *Medical Decision Making,* 4, 47-58.

Dworkin, G. (1983), 'Paternalism' in R. Sartorius (ed), *Paternalism,* Minneapolis: University of Minnesota Press, pp.19-34.

Elster, J. (1979), *Ulysses and the Sirens: Studies in Rationality and Irrationality,* Cambridge, UK: Cambridge University Press.

Feinberg, J. (1986), *Harm to Self: The Moral Limits of the Criminal Law,* New York: Oxford University Press.

Goldman, SM. (1980), 'Consistent Plans', *The Review of Economic Studies,* 47 (3), 533-537.

Gul, F. and W. Pesendorfer (2001), 'Temptation and Self-Control', *Econometrica,* 69 (6), 1403-1435.

Gul, F. and W. Pesendorfer (2004), 'Self-Control and the Theory of Consumption', *Econometrica,* 72 (1), 119-158.

Gul, F. and W. Pesendorfer (2007), 'Harmful Addiction', *The Review of Economic Studies* 74 (1), 147-172.

Hedden, B. (2015) *Reasons without Persons: Rationality, Identity, and Time*, Oxford: Oxford Univesity Press.

Herrnstein, R.J. and D. Prelec (1992),'A Theory of Addiction' in G. Loewenstein and J. Elster (eds),*Choice over Time,* New York: Russell Sage Foundation, pp.331-360.

Holton, R. (2009), *Willing, Wanting, Waiting,* Oxford: Oxford University Press.

Laibson, D. (1997), 'Golden Eggs and Hyperbolic Discounting', *The Quarterly Journal of Economics,* 112 (2), 443-477.

Marneffe, P. de (2006), 'Avoiding Paternalism', *Philosophy and Public Affairs,* 34 (1), 68-94.

O'Donoghue, T. and M. Rabin (1999), 'Doing It Now or Later', *The American Economic Review,* 89 (1), 103-124.

O'Donoghue, T. and M. Rabin (2001), 'Choice and Procrastination', *The Quarterly Journal of Economics,* 116 (1), 121-160.

Peleg, B. and ME. Yaari (1973), 'On the Existence of a Consistent Course of Action when Tastes are Changing', *The Review of Economic Studies* 40 (3), 391-401.

Pollak, RA. (1968), 'Consistent Planning', *The Review of Economic Studies,* 35 (2), 201-208.

Rawls, J. (1971), *A Theory of Justice,* Cambridge, Massachusetts: Harvard University Press.

Salvat, C. (2014), 'Behavioral Paternalism', *Revue de Philosophie Economique/Review of Economic Philosophy,* 1 5 (2), 109-130.

Schelling, TC. (1960), *The strategy of conflict,* Cambridge, Massachusetts: Harvard University Press.

Schelling, TC. (1984), 'Self-Command in Practice, in Policy, and in a Theory of Rational Choice', *The American Economic Review,* 74 (2), 1-11.

Schelling, TC. (1992), 'Self-Command: A New Discipline', in G. Loewenstein and J. Elster (eds), *Choice over Time,* New York: Russell Sage Foundation, pp.167-176.

Strotz, RH. (1955), 'Myopia and Inconsistency in Dynamic Utility Maximization', *The Review of Economic Studies,* 23 (3), 165-180.

Christophe Salvat is researcher at the Centre National de la Recherche Scientifique (CNRS) and member of Triangle, UMR5206, at the Ecole Normale Supérieure de Lyon (christophe.salvat@ens-lyon.fr).