Nudging for changing selves[*]

Richard Pettigrew

September 12, 2021

When is it legitimate for a government to 'nudge' its citizens, in the sense described by Richard Thaler and Cass Sunstein (2008)? In their original work on the topic, Thaler and Sunstein developed the *'as judged by themselves' (or AJBT) test* to answer this question (Thaler & Sunstein, 2008, 5). In a recent paper, L. A. Paul and Sunstein (ms) raised a concern about this test: it often seems to give the wrong answer in cases in which we are nudged to make a decision that leads to what Paul calls a *personally transformative experience*, that is, one that results in our values changing (Paul, 2014). In those cases, the nudgee will judge the nudge to be legitimate after it has taken place, but only because their values have changed as a result of the nudge. In this paper, I take up the challenge of finding an alternative test. I draw on my *aggregate utility account* of how to choose in the face of what Edna Ullmann-Margalit (2006) calls *big decisions*, that is, decisions that lead to these personally transformative experiences (Pettigrew, 2019, Chapters 6 and 7).

1 What are nudges?

Sometimes, your life doesn't go as well as it might because of decisions that you make. You eat unhealthily and get sick; you don't save for your retirement and live in straitened circumstances in your old age; you don't get vaccinated against a dangerous disease and you catch it. Paternalists hold that it is legitimate for governments to intervene to improve your life by ensuring that you don't make such choices. Sometimes, those interventions involve either removing the bad choice or making it extremely onerous to make it. The government might place very high taxes on unhealthy foods, or perhaps ban them outright; they might introduce a mandatory pension scheme to which all employees must contribute; they might require you to

[*] **Acknowledgments** I'd like to thank: Chris Bertram, Martin Sticker, Kevin Backwell, and the audience at Bristol Value Theory Workshop, Sep 2021; Michael Cholbi, Marina Moreno, Adrianno Mannino, Marcus Arvan, and the audience at the MANCEPT 2021 workshop on Prudence and Politics.

be vaccinated against various diseases before you can participate in certain aspects of civic life. The libertarian is horrified. Such restrictions on freedom of choice and autonomy are anathema to them, and well beyond the legitimate reach of government. However, according to some paternalists, there are types of intervention that are likely to improve the lives of those they affect while being perfectly compatible with libertarianism. Enter the nudge theorists, or libertarian paternalists (Thaler & Sunstein, 2003, 2008). According to them, a government can improve its subjects' lives in some of the ways that the paternalist would like it to, but without restricting the freedom of choice that those subjects enjoy and without trespassing on their autonomy.

How? There are three claims that are central to the libertarian paternalist's strategy. First: in many of those cases in which your life goes poorly because of decisions you make, the option you choose is not the best means to your ends. Your end is living the longest and healthiest life that your particular body will afford you, yet you don't choose the foods that are the best means to that end; your end is a good quality of life at all stages of your life, yet you don't save for retirement; your end is a healthy life in the immediate future, yet you choose not to get vaccinated against certain common illnesses.

Second: sometimes when you choose these suboptimal means, you do so because of certain cognitive mechanisms that we all share: perhaps addiction, perhaps a drive towards immediate gratification, perhaps an inability to reason well with probabilities, perhaps a sort of inertia that makes it difficult for us to abandon the status quo even when there is an option available that we prefer. These mechanisms disrupt our rational decision making in some way and lead us to irrational choices that are poor means to our ends.

The third claim that is central to the libertarian paternalist's strategy: in those cases in which certain cognitive mechanisms disrupt our rational decision making and lead us to choose poor means to our ends, there are ways in which the decision might have been presented to us that would lead us to choose the best means. These ways of presenting the choice are called 'nudges'. When we are nudged in these ways, we will most likely choose the best means to our ends, but we might not choose those because they are the best. Indeed, we might end up choosing them for rather poor reasons. For instance, as we'll see below, we might choose to eat a healthy snack simply because it was listed first among the options. But so long as we do choose it, we will have chosen a better means to our ends, and our lives will likely be improved. What's more, since all that the nudge changed was the way in which the decision was presented and not the set of options that were available, the libertarian should be satisfied that freedom of choice, liberty, and autonomy were preserved.

Let's illustrate the strategy with an example. As I've mentioned, many people have the goal of living as long and as healthy a life as their particular body will allow them. They also know that eating certain foods is a good means to that end, while eating others is not. Nonetheless, when standing in the queue at the cafeteria and faced with a choice between the good means and the bad means, people with this end often choose the bad means. Let's suppose that you are a civil servant charged with designing the menu for the cafeteria in some state-run institution, such as a museum. You read the 'first is best' study by social psychologists Dana Carney and Mahzarin Banaji. Here's an excerpt from the abstract of that study:

> We experience the world serially rather than simultaneously. A century of research on human and nonhuman animals has suggested that the first experience in a series of two or more is cognitively privileged. We report three experiments designed to test the effect of first position on implicit preference and choice using targets that range from individual humans and social groups to consumer goods. Experiment 1 demonstrated an implicit preference to buy goods from the first salesperson encountered and to join teams encountered first, even when the difference in encounter is mere seconds. In Experiment 2 the first of two consumer items presented in quick succession was more likely to be chosen. (Carney & Banaji, 2012)

What's more, worried about what you've heard of the replication crisis in social psychology in general and priming research in particular, you look further into this study and find that it seems to replicate. As a result, when you design your menu, you list the healthier options at the top and the less healthy further down. In this way, you hope to sway customers back from the temptation they feel to choose the unhealthy option, which is the poorer means to their end of a long and healthy life, and towards the healthy option. The libertarian is happy because you have not removed any options, but merely presented them in a particular way, and you have not made any of them too onerous to choose. And the paternalist is happy because you have improved the life of the chooser.

Indeed, note something further: you have improved the life of the chooser *by their own lights*. After all, we specified that the chooser has the goal of a long and healthy life. So the nudge you have performed is acceptable to the *means paternalist* as well as the *ends paternalist*. According to the means paternalist, it is legitimate for governments to intervene to make it more likely that citizens will make choices that are better means to the ends they already have; according to the ends paternalist, it is also legitimate to intervene to make it more likely that citizens will make choices that are better means to certain ends that the government takes to be better than the ends the citizens actually have. Nudges can, of course, be used in the service of either form of paternalism. According to the ends paternalist, if the civil servant thinks that few people actually have the goal of a longer and healthier life, but believes that this goal is better than those they actually have, they might appeal to Carney and Banaji's research to nudge people towards healthier eating. However, Thaler and Sunstein are, for the most part, means paternalists, and the test they propose to gauge the legitimacy of a nudge is intended to test whether the nudge is legitimate from that point of view.

2 The 'as judged by themselves' test

We've met nudges now, and we've noted that their enthusiasts are often means paternalists. Let's now meet the test that Thaler and Sunstein suggest we use to identify when a nudge is acceptable to the means paternalist (Thaler & Sunstein, 2008; Sunstein, 2018). As Paul and Sunstein present it:

> The ['as judged by themselves' or] AJBT criterion, as we shall call it, asks whether those who have been nudged *ex ante*—for example, with a warning or a reminder—deem themselves to be better off *ex post* as a result. (Paul & Sunstein, ms, 2)

And here is my paraphrase: a nudge is legitimate if the nudgee would assent to it if asked in a certain idealized situation. What is the idealized situation? It takes place after the nudge has happened; the nudgee is given as much time as they need to reflect on the choice; their irrational cognitive mechanisms, such as status quo bias, temptation, etc. are removed; they are provided with all the evidence relevant to the choice; any cognitive limitations, such as limitations in their logical or statistical reasoning, are removed; and they are equipped with unlimited cognitive resources, such as computing time and power, with which to assess the evidence. This is Thaler and Sunstein's 'as judged by themselves' test.

One distinctive feature of this test is that it is hypothetical. It does not require that the nudgee is *actually* asked the question in the idealized circumstances just described. All that is required is that they *would* give a particular answer *were* they asked it in the those circumstances. After all, we have no way to endow someone with unlimited cognitive facilities; no way to remove all of their cognitive biases; and so on.

Another distinctive feature is that it involves some normative notions—it talks of irrational cognitive mechanisms and cognitive limitations. In order to administer the test in many cases, you will have to settle some controversial normative questions. After all, there is widespread disagreement among economists, philosophers, and psychologists about what count as an irrational cognitive mechanism. Is it irrational to be risk-averse in such a way that you have the Allais preferences when faced with those choices (Allais, 1953)? Is it irrational to be ambiguity averse so that you have the

Ellsberg preferences when faced with those choices (Ellsberg, 1961)? Is it irrational to discount your future selves (Frederick et al., 2002)? If some ways of discounting are rational, which are they? Is it irrational to deviate from the Principle of Indifference when setting your prior probabilities? To each of these questions, there is a reasonable number of incompatible answers each of which is defended by a substantial group of theorists. And for each, there is some nudge that passes Thaler and Sunstein's test when you give one of these answers, but not when you give another.

Let's see this play out in a standard example (Lecouteux, 2015). Suppose I nudge you to make greater contributions to your pension scheme now. You can continue to contribute £140 each month, or your can increase it to £200. Left to your own devices, you are going to choose the status quo; I nudge you to switch to the higher contribution. Is this nudge legitimate? Well, according to Thaler and Sunstein, that depends on whether you'd assent to it in the idealized situation described above. And that will depend on whether, in this idealized situation, you discount the future, and if so, by how much. And that, in its turn, depends on whether discounting the future, or discounting it to the extent you in fact do, is an irrational cognitive mechanism. Suppose you discount the future quite dramatically, and that's why you wish to stick with your current contribution. While the extra £60 will bring you less happiness now, when you are reasonably well off, than it would when you have retired, when you will be much less well off, you discount that future retired self so much that it compensates for this extra happiness. Some will take this dramatic discounting to be irrational, and will expunge it when they move you to the idealized situation in which Thaler and Sunstein's test is administered. Others will not. This leaves us with a question: which theory of rational choice should be used when we apply Thaler and Sunstein's 'as judged by themselves' test? The nudger's or the nudgee's? The means paternalist is happy to override the means we take to our ends, but not to override the ends themselves. What about our theory of which is the best means to our ends? Are they happy to override that? We'll return to this question below.[1]

One final point about Thaler and Sunstein's test before we move on the objection to it raised by L. A. Paul and Cass Sunstein. Some object to nudges on the grounds that many of us value making our choices for ourselves, even if we end up making them poorly as a result. I want to be the author of my own life, they might say, and that includes being the author

[1] While preparing this paper, I was fortunate enough to attend a workshop in which Michael Chobli discussed the fascinating question of what John Rawls, and indeed liberals in general, should say when different members of a society have different accounts of the demands of rational choice: should they find a way to respect that pluralism, just as they respect pluralism about conceptions of the good? Or should they try to impose what they take to be the correct account of rational choice? Since libertarian paternalists are most often liberals, this question is closely related to the question I raise here.

of any mistakes I make. I'd rather choose entirely myself and choose badly than choose under the influence of a nudge and choose well. But Thaler and Sunstein's test can accommodate this. The key lies in the question that we ask the nudgee in the idealised hypothetical situation after the nudge has taken place. If we ask them only whether they are glad *that they made the choice that they did*, we might end up thinking it legitimate to nudge someone even when they don't want to be nudged. After all, such a person will nonetheless be glad *they made the choice they did*, for it has better served their ends. But they will not be glad *they made the choice they did because they were nudged*, for they wish to be the author of their own life. So, in order to ensure we don't nudge the unwilling, we must ask them afterwards whether they are glad they were nudged to make the choice they did, not only whether they are glad they made the choice they did.

3 Paul and Sunstein on the test

In this section, I turn to a different concern about Thaler and Sunstein's test. Start by noting that the test is administered—or, better, it is hypothetically administered—after the nudge has taken place. As Paul & Sunstein (ms) put it above, we ask whether the nudgee "deems themselves to be better off *ex post* as a result [of the nudge]". But, as a number of philosophers, economists, psychologists have noted, our values change across time, and sometimes as a result of choices that we make.[2]

Perhaps the most widely discussed example in the philosophical literature is the decision to become a parent.

> **Happy Parent** I am currently child-free, and I am deciding whether or not to adopt. At the moment, I value remaining child-free more than I value adopting a child and becoming a parent. When I look forward to the two possible futures ahead of me, one in which I am a parent and one in which I am child-free, I value the latter more. I value the things I will be free to do in that possible future: the time it will allow me to strengthen and deepen the bonds with my friends; the volunteering opportunities it will afford me the time to pursue; the extra money I'll have available that I can use to pursue the projects I love and donate to the causes that matter to me; and so on. However, I've spoken to enough parents to know that, were I to adopt, these preferences are likely to reverse. I will likely form a bond with my adopted child so strong that I will prefer the life in which I care for them to the one in which they are not in my life.

[2] Some representative pieces from the philosophical literature: (Parfit, 1984; Ullmann-Margalit, 2006; Bykvist, 2006; Paul, 2014).

And there are other cases as well. Should I emigrate to another country, where the dominant values are different from those in my home country, there is evidence that I will likely change my values to better match those of my new surroundings, at least to some extent (Bardi et al., 2014).

So now imagine that the government were to nudge me towards abandoning my currently preferred child-free life. And suppose they were to succeed: I adopt and become a parent. From the moment my adopted child comes to live with me, my values change in exactly the way I predicted. At this point, *ex post*, we administer Thaler and Sunstein's test. Am I glad I was nudged? Yes! Of course I am! For I currently prefer the life I have to the alternative in which I remained child-free. So the nudge passes the test: it's deemed legitimate. And yet this seems the wrong verdict. It does not seem a legitimate nudge. Not only has the government nudged me to take a means to an end I don't have; but they've done so knowing that, by doing this, I will change my ends as well. The means paternalist is horrified.

4 A natural tweak?

A natural first reaction to this problem is to think that it admits of a straightforward solution. Surely there is an simple tweak to Thaler and Sunstein's test that will allow it to cope with these cases. Instead of administering the hypothetical test only after the nudge, we administer it both before *and* after. A nudge is then deemed legitimate if the nudgee would be happy with it when asked under idealized conditions at both times. Since it seems that I would not be happy with being nudged to adopt when asked before the nudge, that nudge is not legitimate, just as we suspected.

Now, you might worry that Thaler and Sunstein's test must be administered only after the nudge has taken place because nudges often only work if the nudgee is not aware they're being nudged. If you're told that an option has been placed at the top of the list because the person who made the list wants you to choose it, your contrarian side might kick in and you might be minded to thwart their attempt by choosing something else, or you might simply randomise so as not to feel a dupe. But remember that, as we noted above, the test is purely hypothetical—we do not in fact administer it. Rather, the nudger asks themselves what the nudgee would say were they to be asked in the idealized circumstances Thaler and Sunstein describe. So there is no concern about the test interfering with the efficacy of the nudge.

5 Harman's 'I'll be glad I did it reasoning'

Nonetheless, there is a problem. To see it, it's helpful to note that Thaler and Sunstein's original test is essentially a third-person version of the form of reasoning that Elizabeth Harman calls 'I'll be glad I did it' reasoning (Harman, 2009). I find myself on my sofa on a cold winter's night; I haven't really moved all day, and I've got a cup of steaming hot tea and the book I'm enjoying near at hand. Despite this enviable position, and despite the fact I'd currently prefer to stay on the couch, I decide to go for a run. Why? When asked, I justify my choice by saying: If I go for a run, I'll be glad I did it.

As Harman points out, this might seem reasonable in the case just described, but it is not good reasoning in general. Here's an example I have given in which it goes wrong; it is close to one of Harman's original examples (Pettigrew, 2019, Chapter 15):

> **Deborah's pregnancy** Deborah has decided to have a baby, but she needs to decide when to try to become pregnant: now, or in three months' time. Currently, she has a virus, and she knows that, when people become pregnant whilst carrying this virus, their child will have an extremely high chance of developing a very aggressive cancer around the age of forty. However, if she becomes pregnant in three months' time, once her body is rid of the virus, there will be no risk to her child. Currently, she values having the child with the prospect of aggressive cancer very much less than she values having the child without. However, if she becomes pregnant now and has a child with that prospect, she will, most likely, form a bond with them so strong that she would value having that particular child, with their tragic prognosis, more than having any other child, including the child without that prognosis that she would have had if she had waited three months. After all, the alternative child would have been a different child, created from different gametes; they would not be the child with whom Deborah has formed the bond. So, if Deborah becomes pregnant now, she'll be glad she did it. Nonetheless, that seems like a bad reason to do so.

Thaler and Sunstein's test essentially says this: a nudge is legitimate if the nudgee would be glad the nudger did it. And many of the nudges that it incorrectly judges as legitimate are ones that lead to decisions we might try but fail to justify using 'I'll be glad I did it' reasoning. For instance, it says it would be legitimate to nudge Deborah to become pregnant now rather than in three months.

Now, just as we tried tweaking Thaler and Sunstein's test to overcome these problems in the case of nudges, so we might think we can overcome analogous problems here by amending 'I'll be glad I did it' reasoning in the same way. The fact you'll be glad you did something is not, on its own, a good reason to do it; however, we might think that, combined with the fact that you currently want to do it, it is. Again, we suggest that, instead of only asking whether it's what you prefer after the decision, we ask both before and afterwards.

Nonetheless, this tweak still fails. Consider the following sort of case (Bykvist, 2006; Pettigrew, 2019):

> **Unhappy Parent** I am currently child-free, and I am deciding whether or not to adopt. At the moment, I greatly value the prospective future life in which I am a parent. I also greatly value the future life in which I remain child-free, but I value that slightly less. If I remain child-free, I'll retain these values. If I become a parent, I will come to assign a pretty low value to the life of a parent, but I'll assign even lower value to the alternative life in which I'm child-free. Becoming a parent will lower the value I assign to my life more generally, but I will retain the view that this life with a child is more valuable than it would be without.

In this case, it seems, I'll want to become a parent before I do, and I'll be glad I did it afterwards; and indeed, if I remain child-free, I'll regret that, because I'll at that time still prefer the life of a parent. If I am nudged into becoming a parent, I'll be happy with this should I be asked in the idealised situation both before the decision and afterwards. So the nudge counts as legitimate according to the tweaked version of Thaler and Sunstein's test that I described in the previous section. Yet it seems I don't have good reason to become a parent in this case; and it seems illegitimate to nudge me towards that life. It would surely be better for me to remain child-free and live with the slight regret that results from living a life that I slightly disprefer to an alternative I might have lived; better than live a life I would value very little were I to live it.

If this is right, even the tweaked version of Thaler and Sunstein's test fails. How, then should we test the legitimacy of a nudge?

6 Choosing for changing selves

In fact, I think the tweaked version of Thaler and Sunstein's test is on the right track. But, in the apparent counterexamples I have given, I have applied it in the wrong way. When I've applied it in the cases I've described so far in this paper, I've answered the question whether the individual would be happy with the nudge at a particular time by looking to what I have called elsewhere their *local utilities* at that time (Pettigrew, 2019, 18). These are the utilities that encode their values at that time. So, in Unhappy Parent, the local utility I assign to being a parent before becoming one is very high, while the local utility I assign to remaining child-free is still high, but slightly lower; and so on. However, as I've argued elsewhere, you should not use your local utilities when you make a choice at a particular time (Pettigrew, 2019, Chapters 6 and 7). Rather, those local utilities constitute just one factor that goes into determining what I call your *global utilities* at that time, and these are the ones you should use to make your decisions. The other factors that determine your global utilities at a time are your local utilities at other times in your life.

The argument runs as follows. When you make a decision at a particular time, your current self at that time makes the decision on behalf of all the selves that make up the person you are—your past, present, and future selves. Because of this, the utilities you use to make the decision should be the result of aggregating the local utilities of each of those selves. Figuring out how you should aggregate these local utilities to give your global, decision-making utilities at a particular time is akin to figuring out how a state should choose on behalf of a citizenry with a wide variety of values, or how the head of an activist collective should choose on behalf of its membership, many of whom have differing ends. It is the central problem of social choice theory.[3]

The aggregation method for which I argued runs as follows. Let's represent a possible outcome of a choice you make as an entire possible history of the world, which includes: (a) the fixed history up to the moment of the choice and specifies, at each point at which you exist in that history, the local utility that encodes the extent to which you, at that point of time, value the whole history, and (b) the future development of the world that again specifies, at each point at which you exist, your local utility at that point for the whole history. We must then aggregate the various local utilities that you assign to this outcome at the various points at which you exist within it to give your current global utility for that outcome, which you'll use to make decisions now. And we do that by assigning a weight to each of your selves that exist in this history based on certain considerations, weight their utility by that, and then sum up those weighted utilities.

How are we to assign the weights? There are various considerations.

[3] Of course, when I tell you that, in order to solve the problem of rational choice for change selves I must solve the central problem of social choice theory, you might immediately respond that various impossibility theorems from Arrow's onwards show that there will be no satisfactory solution (Arrow, 1951; Gaertner, 2009). In fact, as I argue, this isn't quite true for the particular case that is our focus here (Pettigrew, 2019, Sections 6.3 and 7.3). For one thing, we assume cardinal utilities, and so Arrow's theorem itself does not apply. The results that come closest to threatening my proposal are due to Philippe Mongin (1995) and Matthias Hild (2001). However, I argue that these in fact do not cause any problems. Mongin's does not apply because we aggregate the local utilities of the various selves to give the decision-making utility, but we don't aggregate the credences of the various selves to give the decision-making probabilities—we just use our current credences. And Hild's does not apply because there is a privileged level of grain at which we describe the outcomes.

Some of them impose genuine obligations to assign weights in a particular range; some impose no obligations but are the sorts of considerations that we might adduce to justify the weights that we do assign.

So, for instance, the fact that the other selves form part of the same person as your current self, and the fact that you are choosing on behalf of that person and not just your current self, creates a defeasible obligation to assign at least some weight to the local utilities of each other self. What might defeat this obligation? If one of the selves that belongs to the person you are has local utilities that you take to be morally abhorrent, that would defeat your obligation to give them any weight at all. If my past self valued eating meat, while my current self finds that morally beyond the pale, I need not give that past self's local utilities any weight.

As well as the general defeasible obligation to give each self at least some weight, there are further, more specific obligations to give particular selves greater weight. For instance, if a past self has made a sacrifice from which your current self benefits, you might have an obligation to give that past self some significant weight. Or, if your current decision will disproportionately affect certain future selves, you might have an obligation to give those significant weight.

And then there are considerations that do not create obligations. For instance, we often assign greater weight to selves in whom we recognise ourself more or with whom we anticipate a greater degree of psychological connectedness (Parfit, 1984, 313). So I might give low weight to a future self who is a parent because I find them and their values alien to my current way of thinking. But there is no obligation to do this. It's also open to me, once I have given weight in accordance with the obligations described in the previous two paragraphs, to then divide the weights as equally as possible among all selves.

This overview is inevitably a little brief, but it sketches some of the central theses of my account of how an individual should choose when the selves that constitute them have different values. They should use their global utilities at the time of the choice, which are weighted averages of the local utilities that encode the values of those different selves. My proposal here is that it is also these global utilities, and not the local utilities that partly determine them, to which we should appeal when we administer the tweaked version of Thaler and Sunstein's test. That is, when we ask the nudgee before and after the nudge whether they are glad of it, we are asking whether it is a better means to the ends encoded in their global utilities than the choice they would otherwise have made; we are not asking whether it is a better means to the ends encoded in their local utilities.

7 The new 'as judged by themselves' test at work

To see how this would work, let's apply it to two of the examples we've described above. My description of these is idealized in certain ways in order to make the fundamental idea most apparent.

First, Happy Parent. To simplify greatly, we suppose that, if I apply to adopt, I'll be successful. So there are two available options: *Adopt*, *Don't Adopt*. And there are two possible outcomes: *Become a parent*, which is sure to happen should I choose *Adopt*; and *Remain child-free*, which is sure to happen should I choose *Don't Adopt*. Throughout my life up to the decision point, I have valued the child-free life a great deal, and a little more than the parental life. If I remain child-free, I'll retain those values. If, on the other hand, I become a parent, I'll value the parental life enormously, and the child-free life a lot less. To help us think this through, we might put some numbers on these values—that is, we might measure them numerically as local utilities as follows:

	Child-free	*Parent*
Before	12	8
After & Adopt	4	128
After & Don't Adopt	12	8

Now let's aggregate these local utilities to give my global utilities for the two possible outcomes, *Parent* and *Child-free*. To do this, for each outcome, I need to assign weights to the various selves that make up the person that is me in that possible history; and I need to do this from the point of view of my current self. Let's look first at the outcome *Parent* in which I adopt a child. I consider my local utilities both now and after becoming a parent to be morally acceptable, so I'm obliged to give at least some weight to both current and future selves. But how much weight to each? Here, my obligations end and I am permitted to do a number of things. For instance, just as many think it's legitimate in our interactions with other people to give more weight to ourselves and our nearest and dearest, so it's legitimate for my current self to give more weight to itself than to my future self. And, in particular, my current self might assign quite a lot more weight to itself than to my future self because, since the values of my future self are so dramatically different from those of my current self, my current self does not fully recognise themselves in that future self. Let's suppose my current self actually assigns three times as much weight to itself as to my future self.[4] In this case, we have the following global utility for the outcome

[4] We assume here that weights always sum to 1. In the cases we will consider, in which each possible history contains the same number of selves, this is an innocent assumption. But when different histories contain different numbers of selves, it becomes a strong and possibly implausible assumption for the same reason that average utilitarianism is implausible. Since it won't affect the points I wish to make here, I'll leave the assumption in place. I thank Adrianno Mannino for emphasising the connection with population ethics here.

Parent in which I adopt a child:

$$\begin{aligned}&\text{Global Utility in } \textit{Parent} \text{ before adopting} =\\ &\quad(\text{weight for current self} \times \text{current self's local utility for } \textit{Parent})+\\ &\quad(\text{weight for future self} \times \text{future self's local utility for } \textit{Parent}) =\\ &\qquad\left(\frac{3}{4}\times 8\right)+\left(\frac{1}{4}\times 128\right)=38\end{aligned}$$

And, since your values in the outcome *Child-free* don't change throughout the history that represents that outcome, whatever weights you assign, your global utility for that outcome is:

$$\begin{aligned}&\text{Global Utility in } \textit{Child-free} \text{ before adopting} =\\ &\quad(\text{weight for current self} \times \text{current self's local utility for } \textit{Child-free})+\\ &\quad(\text{weight for future self} \times \text{future self's local utility for } \textit{Child-free}) = 12\end{aligned}$$

So, although your current local utility for becoming a parent is lower than your current local utility for remaining child-free, your global utilities at the earlier time, which incorporates the local utilities of your current and future selves, are ordered the other way around. That is:

$$\begin{aligned}&\text{Global Utility in } \textit{Parent} \text{ before adopting} >\\ &\qquad\text{Global Utility in } \textit{Child-free} \text{ before adopting}\end{aligned}$$

What's more, at the later time, after the decision is made, if you chose to be a parent then, providing you will give more weight to your current self at that time than to your past self, then:

$$\begin{aligned}&\text{Global Utility in } \textit{Parent} \text{ after adopting} >\\ &\qquad\text{Global Utility in } \textit{Child-free} \text{ after adopting}\end{aligned}$$

In this case, then, it would be legitimate for the government to nudge me to adopt. After all, when we look at my global, decision-making utilities before and after the nudge, I'd be glad of the nudge.

On the other hand, here is the situation described in Unhappy Parent, with some indicative numbers used:

	Child-free	*Parent*
Before	120	128
After & Adopt	8	12
After & Don't Adopt	120	128

Then, using the same weights that we used above:

$$\begin{aligned}&\text{Global Utility in } \textit{Parent} \text{ before adopting} = \\ &\quad (\text{weight for current self} \times \text{current self's local utility for } \textit{Parent}) + \\ &\quad (\text{weight for future self} \times \text{future self's local utility for } \textit{Parent}) = \\ &\qquad \left(\frac{3}{4} \times 128\right) + \left(\frac{1}{4} \times 12\right) = 99\end{aligned}$$

And, since your values in the outcome *Child-free* don't change throughout the history represents that outcome, your global utilithy for that is

$$\begin{aligned}&\text{Global Utility in } \textit{Child-free} \text{ before adopting} = \\ &\quad (\text{weight for current self} \times \text{current self's local utility for } \textit{Child-free}) + \\ &\quad (\text{weight for future self} \times \text{future self's local utility for } \textit{Child-free}) = 120\end{aligned}$$

So, in this case,

$$\begin{aligned}&\text{Global Utility in } \textit{Parent} \text{ before adopting} < \\ &\qquad \text{Global Utility in } \textit{Child-free} \text{ before adopting}\end{aligned}$$

And if the government nudges me to adopt in the Unhappy Parent scenario, the tweaked version of Thaler and Sunstein's test that I propose deems it illegitimate, just as we would like.

So this is our proposed test: a nudge is legitimate if the nudgee's global utilities before the nudge lead to preferences that favour it and the nudgee's global utilities after the nudge do likewise.

8 The problem of the weights

Now, at first sight, this might seem a rather different sort of test from the one that Thaler and Sunstein describe. Mine talks of actual global utilities, which I take to represent actual internal mental states, even if somewhat idealised, while theirs talks of choices in idealised hypothetical situations, which are hypothetical external behaviours. But in fact I think both seek to pinpoint the same thing. I take it that Thaler and Sunstein assume that, in the hypothetical situation they describe, what you choose is what it would be rational for you to choose given your true preferences, your true utilities, your true credences, your true attitudes to risk, and so on.[5] If not,

[5] Infante et al. (2016) agree with this interpretation of Thaler and Sunstein's intention—what I am calling the true preferences, utilities, credences, and atittudes to risk, Infante, et al. call the attitudes of the "inner rational agent". They raise worries about the psychological reality that this picture assumes. I don't.

why move to this hypothetical scenario? As I noted above, one of the central tenets of nudge theory is that some of our cognitive mechanisms and limitations lead us to choose suboptimal means to our ends. As a result, nudge theorists hold that looking to our actual choices won't reveal our true preferences or our true utilities in the way that economists have sometimes assumed. Instead we must look to our hypothetical choices in the idealized situation in which these cognitive mechanisms and limitations are removed. In those situations, our true utilities are revealed. After all, when nudge theorists say that we often take suboptimal means to our ends, they are thereby assuming that we have ends, even if they are sometimes obscured by our cognitive mechanisms and limitations. And it is these ends that are encoded in our local utilities and then aggregated to give our global utilities.

Nonetheless, it is true that my version of Thaler and Sunstein's test requires something more than merely our ends, which are encoded in our local utilities. It also requires the weights we apply to these local utilities when we generate the global, decision-making utilities to which we appeal in the revised version of the test. This raises two problems: first, these weights are often much more difficult to discover than the local utilities; second, often, the nudgee does not set these weights in advance of making the decision for which they are required. Let's treat these two issues in order.

8.1 Unknown weights

First, suppose I have set the weights I will apply to my own local utilities and to those of the other selves in the collective that makes up the person I am. How might you, as a prospective nudger keen to figure out whether your nudge would be legitimate, go about discovering these? There seem to be (at least) four sources of information on which you might draw: my testimony and my past choice behaviour; and the testimony and choice behaviour of others. We talk about the weights we'll assign to our future and past selves less often than we talk about our values, goals, and ends, but we aren't completely silent about them. For instance, by listening to the testimony of others and observing the choices they make, you might note that some people assign lower weight to future selves the less they identify with them. Bearing this in mind, and observing my past choices, you might notice a close relationship between how much I say I identify with a future self and the weights I must be assigning to them in order to justify the choices I make. So, if I say before adoption that I simply don't recognise myself in the person I think I'll become after adopting, that's good evidence that I'm going to assign them very low weight. So, while it might require greater effort to gather the sort of evidence we need to discover that most people assign less weight to future selves with whom they identify less than to gather the evidence we need to discover that most people value lives more the longer and healthier they are, it is nonetheless possible.

It's worth noting again before we move on that this is no purely theoretical puzzle. As we've already seen, governments will have to know the weights that individuals assign to their future selves in order to assess the legitimacy of some of the most standard nudges. After all, whether or not it is legitimate to nudge someone to contribute more to their pension depends in part on the extent to which they discount the utilities of their future selves. And that is essentially the question of how much weight they assign to those future selves.

8.2 Undetermined weights

Let's now turn to the second problem: sometimes the weights my current self assigns to past, present, and future selves' local utilities to give my current global utilities just don't exist yet; sometimes, I just haven't set these weights. In such a case, is it legitimate to nudge me in one direction or another?

In fact, once again, this question arises in the case of the nudge towards pension contributions, since many people simply haven't thought carefully about the extent to which they discount the future and therefore haven't set their discount rate. So the question arises: in such a case, is it legitimate to nudge them? Indeed, this may well be the situation for many nudges. It's notable that, in Carney and Banaji's 'first is best' study from above, it is when individuals have no strong prior preferences between the options—the salesperson from whom they buy or the candidate for whom they vote—that they choose the first they encounter. So, if a nudging strategy that appeals to the 'first is best' results turns out to be effective, that suggests that the individual has no strong preferences prior to the nudge, and that might be because they have considered the outcomes and settled on roughly equal utilities for each, or it might be because they have not considered the outcomes and so haven't yet set their utilities for them.

So: is it legitimate to nudge someone into a choice when they do not have set preferences between the options, either because they have not set their local utilities in the options or because they have not set the weights they will apply to those local utilities to give the global utilities they will use for decision-making?

The first and rather predictably philosophical thing to say is that it depends. It depends on what sort of paternalist you are; and indeed thinking about these sorts of cases leads us to draw further distinctions between different varieties of paternalisms. An ends paternalist clearly thinks that it is sometimes legitimate to do this, since they think it's legitimate for the government to nudge you towards a choice that is the optimal means to some ends other than your own and a suboptimal means to your own ends. But what of the means paternalist, which is the brand of paternalism most often associated with nudge theory? Here, I think there are at least two camps. There are those who think it is only reasonable to intervene to make it more likely someone chooses a better means to ends *to which they are currently committed*. We might call these *means-to-existing-ends paternalists*. And there are those who think it's legitimate to intervene to secure the best means to an end *that the intervener determines themselves, but only when the target of the intervention has no current commitments either way regarding that end*. We might call these *means-to-unset-ends paternalists*. And indeed, you might divide the second view into two further positions. On the one hand, there are the *means-to-unset-but-determined-ends paternalists*, who say that it's legitimate to intervene to influence a choice even when the individual has not set their utilities in a particular way, providing the individual *would* set them in a particular way *were* they given the chance to do so, and your intervention points them towards the best means to those ends they would have. On the other hand, there are the *means-to-unset-and-undetermined-ends paternalists*, who says that this is not necessary.

To which of these positions should the nudge theorist or libertarian paternalist subscribe? I'm not sure this question admits of a determinate answer, since there are many different nudge theorists and the commitments they all share might not determine a single answer. But let me consider a couple of the options.

First, consider the *means-to-unset-and-undetermined-ends paternalists*. And think again of the example of Happy Parent. I'm child-free, and I currently prefer that; I'll retain that preference if I remain child-free; but I'll come to vastly prefer being a parent if I choose to do that. My global, decision-making utilities for the two options at the point of decision depend on the weights I assign to my current and future selves. But let's suppose I haven't set them. So, at the moment, I have local utilities but no global utilities. If I were given the chance to set these weights and thereby set my global utilities, I'd do so in a way that favours remaining child-free. Is it legitimate for the government to nudge me in the opposite direction, that is, to become a parent?

A natural answer is that it is not, but for reasons we more often associate with liberalism in general rather than means paternalism specifically. After all, liberalism requires that the government not trespass on my autonomy (unless by exercising my autonomy I trespass on someone else's). And on many accounts of autonomy, what is important is that I am the only one who chooses the way I want to live my life, and that includes the ends that I have and pursue during that life (Raz, 1988; Colburn, 2010). If the government nudges me to become a parent when that is not the option to which I was committed to assigning a higher global utility beforehand, and if by becoming a parent, I come to consider being a parent to be one of my ends, it was to some extent the government and not me who was the author of my ends.

Of course, as is often pointed out, most contemporary liberals recognise that, *pace* Kant, we cannot hope to be the sole author of our ends. Even if we explicitly choose some of our ends, we do so from the starting point of other, perhaps second-order, ends that we have. And there must be some point at which the ends on which we base our choice of other ends are not chosen and come from outside ourselves—from the society we live in, the media we consume, the family we grow up in, the group of friends or colleagues with whom we share so much of our lives, and so on. Many liberals react to this by saying that it is not necessary for our autonomy that we should be the sole author of all of our ends; instead, what is required is that we endorse the ends we have when we reflect upon them and upon the way they were formed (Dworkin, 1976, 1988). But if that is our criterion, then nudging me to become a parent does not trespass on my autonomy. For afterwards and upon reflection, I do endorse the ends that I have come to have as a result of the choice I was nudged to make.

So, if means-to-unset-and-undetermined-ends paternalism is wrong, it isn't because it trespasses on autonomy. But it seems wrong nonetheless. I think a better reason to reject it is the threat of governmental overreach. That is, the problem is not that I was not the sole author of the ends I came to have after the nudge; the problem is that the other actor responsible for those ends was specifically the government. If nudges of the sort we are considering were legitimate, they would provide governments with a legitimate means by which to shape the ends of their citizens. And this, both liberal and libertarian agree, is beyond the pale. Such nudges are wrong for exactly the reason that government-mandated party political propaganda in schools would be wrong. In both cases, the government abuses its power to shape the preferences of its citizens.

Does the government also abuse its power if it nudges you towards the best means to the ends that you *would* set if you *were* to consider them? In other words, does the objection just raised against means-to-unset-and-undetermined-ends paternalism also tell against means-to-unset-but-determined-ends paternalism? I would say not. Requiring that this counterfactual is true puts a strong limitation on government overreach. They cannot shape the preferences of citizens in ways the citizens themselves would not shape those preferences themselves.

Nonetheless, you might think that there is a problem here. It lies in the sort of evidence that the government might gather to justify such a nudge. It will almost certainly be statistical. It will most likely pick out certain features you have, and then note that among people with those features, almost everyone who has considered their ends has the ends that the government will assume you would have were you to consider them. I think some might object to this on the grounds that it treats us not as free individuals who freely choose our ends, but as people whose ends are determined by certain of their features.

I don't find this objection compelling myself. Were the sort of regularity in question observed, it simply would suggest that people with the features in question do tend to choose them in a particular way. And that suggests it will be true of you as well. But it says nothing about why. I do not deny your autonomy if I assume of you, based on my experience of other people, that, if you are presented with the choice between an hour of pleasure and an hour of pain, you'll choose the pleasure. You and everyone else I've observed has been perfectly free to choose either option. But observing that nearly everyone chooses the pleasure gives me very strong reason to think that you will as well. Sometimes there are simply good reasons to do one thing rather than another, and in those cases, many people will do that thing; but this does not mean that they were not free to do otherwise.

9 Disagreements about rationality

Let me close by returning to a question I raised in passing at the end of Section 2. When we administer Thaler and Sunstein's 'as judged by themselves' test, we envisage an idealised situation in which certain features of the prospective nudgee's cognition have been changed, and we ask them whether they'd be glad of the nudge. In particular, in this idealised situation, we remove any irrational cognitive mechanisms at work in the nudgee. This means that, to administer the test, we must know which mechanisms count as irrational and which do not. But it seems like that, just as economists, philosophers, and psychologists disagree over the boundaries of the rational, so might the nudger and the nudgee—indeed, either nudger or nudgee might be an economist, philosopher, or psychologist. In such cases, whose conception of rationality should we use? It's even possible that your conception of rationality might change as a result of the choice you make, and so your current self might disagree with your future self about what counts as rational. Again: to which should we appeal when we construct the idealised situation in which to run Thaler and Sunstein's test, or my revised version of it?

Suppose, for instance, that I discount my future retired self so heavily that I am not now prepared to sacrifice even a modest extra amount of my current income in order to better fund a pension scheme that will benefit that future self considerably. You, a prospective government nudger, think this level of discounting is irrational, whereas I think it's rational. Would it then be legitimate for you to nudge me into saving more for my pension?

Here's one situation in which it would. Suppose I conceive of rationality the way I do because of misleading evidence. Perhaps someone who convinced me that they had read the relevant literature told me that it concludes that this sort of discounting is rationally permissible, and I decide to defer to experts on this matter. And suppose further that, were I exposed to the various arguments that theorists of rationality have actually made, I'd conclude that it was in fact irrational. Then it seems that the nudge would be legitimate, for in Thaler and Sunstein's test, when we construct the idealised situation, we provide the nudgee with all the relevant evidence, and in this case doing that would lead me to change my mind about rationality, and presumably I'd therefore come to favour saving more.

However, many cases that are not like that. Perhaps both nudger and nudgee are theorists of rationality who are aware of all the existing arguments, but simply disagree on the correct conclusion to draw from them. In this case, I think the libertarian paternalist must respect the nudgee's account of rationality and refrain from nudging them. I think this follows from the liberalism that underpins the means paternalism component of nudge theory. Such liberalism seeks to preserve as much diversity of viewpoint as possible, and it is reasonable to include accounts of rationality among those viewpoints.

You might feel that there is a stronger case for nudging someone with a different account of rationality from yours if you were to think that what they count as rational and you count as irrational is not only irrational but also, in some sense, immoral. After all, it's a peculiar feature of conceiving of the person you are as a collective of selves at different times that questions about how you should choose when your choice will affect future selves seems, on the one hand, a question of prudential rationality, since you are asking what will most likely lead to the best life for the person that you are, but on the other hand, a question of morality, since you are asking what will lead to the best outcomes for the different selves you will affect. I think we see some of this dual aspect when we consider the pension case or cases of unhealthy eating. When someone fails to save for their pension or eats unhealthily, the reactions they receive are often closer to moral judgments than to criticisms of prudential rationality. It is as if people criticise the current self of the person who doesn't save or who eats unhealthily on the grounds that they are selfishness or show callous disregard for their future selves, both of which are moral criticisms not prudential ones.

If we do consider failures to give due weight to our future selves to be at least partially moral failings, or something akin to that, does that give us greater reason to nudge someone away from acting on those weights? I think not. Typically, nudge theorists do not advocate using nudges to improve the moral behaviour of the nudgees. They don't suggest nudging people to refrain from romantic infidelity, nor from lying to friends. The reason is that there are certain aspects of morality that we take not to fall under the purview of government. Which aspects we do consider as falling under their purview changes from time to time and place to place, of course, but there will nearly always be some immoral acts that the government has no remit to punish or prevent. If the cases of saving for the future and eating healthily do not fall under the government's purview except insofar as they are ends that the individuals already have, then there is no reason to nudge individuals against the dictates of their faulty account of rationality. On the other hand, if we do consider those cases to fall within the government's remit, we should not use nudges to enforce the prudentially rational or morally required choice, but rather legislation, such as mandatory pension schemes or bans on certain unhealthy foods.

So, in the end, I think it will be rare that we should test for the legitimacy of a nudge by constructing the idealised scenario in Thaler and Sunstein's test using the nudger's conception of rationality, and not the nudgee's.

10 Conclusion

In sum: I think it is sometimes legitimate for the government to nudge people to make choices that they know will result in personally transformative experiences and subsequent changes in values. But the bar the nudger must clear is high. If the nudgee has already set the weights they apply to the local utilities of their various selves both before and after the nudge, the nudger must discover those and ensure that the nudgee would assent to the nudge at both times based on the global utilities obtained from the local utilities using those weights. And if the nudgee has not set those weights, the nudger must have strong evidence that the nudgee would set them in a way that would lead them to assent to the nudge both before and afterwards. In both cases, the knowledge required is hard to come by.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica*, *21*(4), 503–546.

Arrow, K. J. (1951). *Social Choice and Individual Values*. New York: Wiley.

Bardi, A., Buchanan, K. E., Goodwin, R., Slabu, L., & Robinson, M. (2014). Value stability and change during self-chosen life transitions: Self-selection versus socialization effects. *Journal of Personality and Social Psychology*, *106*(1), 131–146.

Bykvist, K. (2006). Prudence for changing selves. *Utilitas*, *18*(3), 264–283.

Carney, D. R., & Banaji, M. R. (2012). First is Best. *PLoS ONE*, *7*(6), e35088.

Colburn, B. (2010). *Autonomy and Liberalism*. London: Routledge.

Dworkin, G. (1976). Autonomy and Behaviour Control. *The Hastings Center Report*, *6*, 23–28.

Dworkin, G. (1988). *The Theory and Practice of Autonomy*. Cambridge, UK: Cambridge University Press.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, *75*(4), 643–69.

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

Gaertner, W. (2009). *A Primer in Social Choice Theory*. Oxford: Oxford University Press.

Harman, E. (2009). 'I'll be glad I did it' reasoning and the significance of future desires. *Philosophical Perspectives*, *23*(1), 177–189.

Hild, M. (2001). Stable Aggregation of Preferences. Social science working paper 1112, California Institute of Technology.

Infante, G., Lecouteux, G., & Sugden, R. (2016). Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, *23*, 1–25.

Lecouteux, G. (2015). In Search of Lost Nudges. *Review of Philosophy and Psychology*, *6*, 397–408.

Mongin, P. (1995). Consistent Bayesian Aggregation. *Journal of Economic Theory*, *66*(2), 313–351.

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

Paul, L. A. (2014). *Transformative Experience*. Oxford: Oxford University Press.

Paul, L. A., & Sunstein, C. (ms). "As Judged By Themselves": Transformative Experiences and Endogenous Preferences. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3455421.

Pettigrew, R. (2019). *Choosing for Changing Selves*. Oxford, UK: Oxford University Press.

Raz, J. (1988). *The Morality of Freedom*. Oxford: Oxford University Press.

Sunstein, C. R. (2018). "Better off, as judged by themselves": a comment on evaluating nudges. *International Review of Economics*, *65*, 1–8.

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian Paternalism. *American Economic Review*, *93*(2), 175–79.

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

Ullmann-Margalit, E. (2006). Big Decisions: Opting, Converting, Drifting. *Royal Institute of Philosophy Supplement*, *81*(58), 157–172.