THEORY AND DECISION LIBRARY

General Editor: Julian Nida-Rümelin (*Universität München*)

Series A: Philosophy and Methodology of the Social Sciences

Series B: Mathematical and Statistical Methods

Series C: Game Theory, Mathematical Programming and Operations Research

SERIES A: PHILOSOPHY AND METHODOLOGY
OF THE SOCIAL SCIENCES

VOLUME 42

Assistant Editor: Martin Rechenauer (*Universität München*)

Editorial Board: Raymond Boudon (*Paris*), Mario Bunge (*Montréal*), Isaac Levi (*New York*), Richard V. Mattessich (*Vancouver*), Bertrand Munier (*Cachan*), Amartya K. Sen (*Cambridge*), Brian Skyrms (*Irvine*), Wolfgang Spohn (*Konstanz*)

Scope: This series deals with the foundations, the general methodology and the criteria, goals and purpose of the social sciences. The emphasis in the Series A will be on well-argued, thoroughly analytical rather than advanced mathematical treatments. In this context, particular attention will be paid to game and decision theory and general philosophical topics from mathematics, psychology and economics, such as game theory, voting and welfare theory, with applications to political science, sociology, law and ethics.

For other titles published in this series, go to
http://www.springer.com/series/6616

PREFERENCE CHANGE
Approaches from Philosophy, Economics and Psychology

Edited by

TILL GRÜNE-YANOFF and SVEN OVE HANSSON
Collegium of Advanced Studies, Helsinki
Royal Institute of Technology, Stockholm

Springer

Rabinowicz, W. and Strömberg, B. 1996. What if I were in his shoes? On Hare's argument for preference utilitarianism. *Theoria* 62: 95–123.
Schueler, G. F. 1984. Some reasoning about preferences. *Ethics* 95: 78–80.
van Fraassen, B. 1984. Belief and the will. *Journal of Philosophy* 81: 235–256.
Vendler, Z. 1988. Changing places. *In Hare and critics: essays on moral thinking*, ed. D. Seanor and N. Fotion, 171–184. Oxford: Clarendon.

Chapter 10
The Ethics of *Nudge*[1]

Luc Bovens

Abstract In their recently published book *Nudge* (2008) Richard H. Thaler and Cass R. Sunstein (T&S) defend a position labelled as 'libertarian paternalism'. Their thinking appeals to both the right and the left of the political spectrum, as evidenced by the bedfellows they keep on either side of the Atlantic. In the US, they have advised Barack Obama, while, in the UK, they were welcomed with open arms by the David Cameron's camp (Chakrabortty 2008). I will consider the following questions. What is *Nudge*? How is it different from social advertisement? Does *Nudge* induce genuine preference change? Does *Nudge* build moral character? Is there a moral difference between the use of *Nudge* as opposed to subliminal images to reach policy objectives? And what are the moral constraints on *Nudge*?

10.1 The Paradigm Cases

I take *Cafeteria* and *Save More Tomorrow* to be paradigm cases of what constitutes a T&S-style *Nudge*, as these cases are repeatedly discussed in their writings (Sunstein and Thaler 2003, pp. 1159–1160 and pp. 1164–1166; T&S 2003, p. 175 and p. 177; T&S 2008, pp. 1–3 and pp. 112–115).

In *Cafeteria*, the goal is to induce students to choose a healthier diet. Studies show that individuals are prone to select items placed earlier and at eye level in a line of food items. So the school's management might try to affect students' diets by rearranging the display of the food items so as to make it more likely that the healthy food items are selected.

L. Bovens
London School of Economics and Political Science, Department of Philosophy, Logic and Scientific Method
e-mail: L.Bovens@LSE.ac.uk

[1] I am grateful for helpful comments to the audiences of the Models of Preference Change Workshop at the GAP6 and of the Choice group in the LSE; to the editors Till Gruene-Yanoff and Sven Ove Hansson; and to Foad Dizadji-Bahmani, Alice Obrecht, Adam Oliver, Esha Senchaudhuri, Peter Sozou, and Alex Voorhoeve.

In *Save More Tomorrow*, the goal is to make employees invest more in their pension fund savings. Employees are asked well ahead of time whether they are willing to commit next year's raise towards their pension funds. They are much more willing to agree to this than when they are asked to do so *after* they have received pay checks with the raises included. This exploits two psychological mechanisms. First, there is the *Endowment Effect*. People tend to find it much harder to part with something once they have it in hand than to forego it when they have never had it in hand. A description of this psychological mechanism can be found in David Hume's *Treatise*[2] and Adam Smith's *The Theory of Moral Sentiments*.[3] Second, people tend to find it much easier to show strength of will when it comes to future than to present costs and benefits. Think of Augustine's Prayer – *make me chaste, but not yet*. What explains the greater willingness to commit to a future loss of income to savings rather than a present loss is a combination of both of these psychological mechanisms.

Nudge is replete with examples that have a structure that is similar to these two cases. What these examples have in common is a manipulation of people's choices via the choice architecture, i.e. the way in which the choices are presented to them. This works in the following way. Choices are structured such that some psychological mechanism leads people toward options that are either thought to be in their own best interest or thought to be in society's best interest. In all cases of *Nudge*, if the choice situation had not been so structured, then people would be less prone to make the choice that is either in their own or in society's interest.

10.2 Social Advertisement

There is a more familiar type of intervention that the government employs to affect our behaviour. In social advertisement campaigns, we are made aware of the dangers of drug usage, the problem of domestic violence, the threat of AIDS, etc. How are such campaigns different from *Nudge*?

Social advertisement affects our choices by providing us with information or by affecting our emotions. Sometimes we learn things that we did not know before and change our behaviour. For example, an addict may change her drug habits after being informed of the death rate associated with cocaine usage. Other times there is no new information offered, but the situation is presented with such force that we change our behaviour. For example, pictures of domestic abuse may induce a wife beater to seek professional help.

T&S do discuss cases of framing in social advertisement (2008, pp. 180–182). For example, social advertisement that conveys the percentage of people who are registered as organ donors is more effective than if it were to convey the percentage

[2] 'Men generally fix their affections more on what they are possess'd of, than on what they never enjoyed (...)' (Hume 1978, Bk III, Part II, Section I; p. 482).

[3] 'To be deprived of that which we are possessed of, is a greater evil than to be disappointed of what we only have an expectation' (Smith 1968, Part II, Section II, Chapter II, p. 94).

of people who are not registered. The information provided is the same, but people are more likely to change their behaviour when the information is positively framed.[4] So social advertisement can be a form of *Nudge*, but not all social advertisement is *Nudge*. So what makes *Nudge* different?

10.3 Rationality and Autonomy

T&S write that their 'basic source of information' is 'the emerging science of choice, consisting of careful research by social scientists over the past 4 decades... [that] has raised serious questions about the rationality of many judgments and decisions that people make' (2008, p. 7). So one defining characteristic of *Nudge*, as opposed to social advertisement that does not qualify as *Nudge*, could be that some pattern of irrationality is being exploited. The psychological mechanisms that are exploited in *Cafeteria* and in *Save More Tomorrow* typically work better in the dark. If we tell students that the order of the food in the *Cafeteria* is rearranged for dietary purposes, then the intervention may be less successful. If we explain the endowment effect to employees, they may be less inclined to *Save More Tomorrow*. And even if we try to affect our own behaviour by means of these mechanisms, then our efforts will be most effective when our knowledge of having done so is latent (or when we simply are able to forget).

The following oft-cited example illustrates this well. People are prone to add an expensive car radio to their newly bought car. But if the car radio is not available on the day of the purchase and they are offered the very same car radio the very next day, then they would never dream of spending this kind of money on a car radio (Savage 1954, p. 103). Now once you point this out to them, they typically try to self-correct. They refrain from buying the expensive radio at the earlier point of time. Or, they may take this to be an argument for spending the money the next day – they remind themselves that they were perfectly happy to buy the radio on the day of purchase. It is not clear what direction they will take the argument, but at least, they will strive for less inconsistency in their actions.

There is something less than fully autonomous about the patterns of decision-making that *Nudge* taps into. When we are subject to the mechanisms that are studied in 'the science of choice', then we are not fully in control of our actions.[5]

[4] An alternative way of distinguishing *Nudge* from social advertisement is to stipulate that a *Nudge* must affect the actual choice situation. So a billboard with cancerous lungs is not a *Nudge*, but a pack of cigarettes with cancerous lungs is a *Nudge*. (I owe this suggestion to Alice Obrecht.) Or we could stipulate it as an additional condition on a *Nudge*. This would be in line with many of T&S's examples, but their example of social advertisement in support of organ donations that appeals to the framing effect would no longer qualify as a *Nudge*.

[5] We may of course use such patterns in an autonomous manner to steer our own agency – as when I rearrange the fridge myself when I am on a diet. But that does not make the action itself of picking the carrots placed in front an autonomous action. My agency is caused by processes that do not constitute reasons. In my quest for weight loss, I can autonomously set up a choice architecture that

When I am presented with full knowledge, then I tend to self-correct my agency. It seems that I was acting on a rule with which I cannot identify. What is so special about the first available item that I would favour it over later items? What is so special about having something in hand that would make it so much more valuable compared to the moment before I have it in hand? Clearly these are cases of not letting my actions be guided by principles that I can underwrite. And in as much, these actions are non-autonomous. Can they be said to be irrational? They can in so far as what is driving my action does not constitute a reason for my action – i.e. it is not a feature of the action that I endorse as a feature that makes the action desirable.

This brings us to the question with which we ended the last section. Why is at least some social advertisement different from *Nudge*? When social advertisement provides us with information that gives us a reason to change our behaviour then the intended effect is again fully autonomous decision-making. If it does not provide us with new information, but increases the saliency of certain reasons then the intended effect is again fully autonomous decision-making. And in this respect there is no reliance on the science of choice that raises questions about the rationality of our decision-making. Of course, these kinds of distinctions are much less clear in the real world. If a social advertiser frames the information in a particular manner that is known to have a greater impact on our decision-making – the more so if we are not made cognisant hereof – then we bring in elements of *Nudge* again. Reasons in support of (or against) the targeted agency and the causal mechanisms that raise (or diminish) the occurrence of the targeted agency mix together in social advertisement; in so far as social advertisement relies on the latter it has a bit of *Nudge* in it.

10.4 What Type of Agency Does *Nudge* Aim to Correct?

I will distinguish between six types of agency that can be made the subject of a *Nudge*.

(i) ***Ignorance***. If the government sets a default for retirement plans, they may do so for the same reason that a medical doctor might recommend a treatment. We typically have little knowledge of the matter at hand. We have a clear goal, viz. to be well off in old age or to recover from an ailment. But the route to this goal requires special expertise, which we lack. So it is lack of knowledge that hampers us in laying out the steps towards realising our goals.

(ii) ***Inertia***. It may be the case that we do have sufficient knowledge, but somehow inertia gets the best of us. We are absorbed in our daily activities and simply put off filling out the forms until we forget. In this case a default option kicking in for the lazy or forgetful may be welcome.

will induce me to act non-autonomously. Consider the following analogy. If I am prone to squeeze a stone in my pocket for good luck, then it may be fully rational to do so when I need to work up the self-confidence in, say, an interview situation. But this does not make the action of squeezing the stone itself rational (cf. Bovens 1995, p. 824).

(iii) ***Akrasia***. Consider our paradigm cases again. Typically we know quite well that our consumption of cream puffs is not conducive towards overall health. We know quite well that we are putting too little money into our retirement funds. What stands in the way is weakness of the will (*akrasia*). We are weak-willed in choosing the proper steps towards our long-term goals. By structuring the choice situation it becomes easier to correct for such weak-willed actions, because temptation will have less of a pull on us.

(iv) ***Queasiness***. In *post-mortem* organ donations, the culprit is not lack of knowledge or weakness of the will. Many of us have no objection to becoming organ donors. It may be inertia, but it could also be queasiness, which prevents us from becoming donors. We are perfectly fine with our organs being used *post-mortem* for transplantation, but we do not want to entertain such matters in decision-making. There is an emotional cost in making the decision of becoming a *post-mortem* organ donor.

(v) ***Exception***. Suppose that particular choices by people with a particular profile tend to engender feelings of regret, whereas alternative choices tend to induce greater *ex post* satisfaction. Let us suppose that there is ample evidence for this in empirical research. For example, we could think of sex change surgeries, abortions, divorces, teenage sex, or what have you. It may well be the case that it holds true as a statistical claim that people who choose some such options typically experience feelings of regret afterwards. Of course there is a reference class problem. It may be the case that for the subgroup to which I belong, suitably defined, this tendency is false. For example, although most transgendered people experience regret after a sex change surgery, the subgroup of transgendered people of which I am a member (say, female, engaged in a relationship with an accepting partner, ...) does not. But suppose that for the narrowest social group for whom we can obtain meaningful results and of which I am a member, this tendency holds. Then I could still claim that, though most people display such feelings of regrets, I, for one, am confident that I will not. And I may be correct in my claim.

(vi) ***Social Benefits***. It may well be the case that a particular individual choice is not socially beneficial. There are many such examples. I may see no benefit whatsoever in giving to charity. In a Tragedy of the Commons, society may be better off if I decide to refrain from, say, adding a fishing boat to over-fished waters or drilling for oil in an oil well that is already quite depleted. But unless I am being compensated or find alternative employment, I may be worse off. In a standard Prisoner's Dilemma, society would have been better off if we had all cooperated, but I would have been worse off if I had cooperated rather than defected. In all such cases, a bit of *Nudge* might be meaningful to realise a socially beneficial outcome, but it may well be at the cost of my own welfare.

In the real world, these distinctions are less pronounced and there are many grey cases. Some cases of *Inertia* may be instances of convenient forgetfulness that are not altogether different from *Akrasia*. *Ignorance* may be intentional because we wish to forego making hard decisions and in this respect such cases may not be altogether different from *Inertia* and *Queasiness*. And I may simply adjust my overall

preferences when I come to learn about the statistical evidence or social benefits and then the only thing that would block my ability to act in a particular case is *Akrasia*. But the existence of these mixed cases does not invalidate the exercise of distinguishing between ideal cases, which will prove useful when thinking about preference change and the moral permissibility of *Nudge*.

10.5 Preference Change[6]

Let us start with *Exception* and *Social Benefits*. In these cases, I am being *Nudged* in the direction of agency which I do not believe to be in my interest. For instance, suppose I believe upon reflection that there is no need to increase my pension savings, but the *Save-More-Tomorrow Nudge* did induce me to do so. What can we say about my preferences when I decide to invest my future raise into my pension fund whereas I would not have done so after the raise had been in place? Did I undergo a preference change?

In one respect the answer is yes – I revealed my preference through my choice. What has changed is that I have a preference for dedicating a greater percentage of my income to my pension fund here and now, which I never had before. In another respect the answer is no. Have I become a more frugal person? Not really. In a way my action is aberrant. It is not well integrated with my overall preference structure – i.e. with my conception of the good, with what I take to be good for me all things considered. I am like the fox and the sour grapes. The fox loses his appetite[7] for the grapes that he cannot reach. Even if he does not want these grapes anymore, he remains the kind of fox who likes juicy summer fruits in general. So his preference over the token action of eating these very grapes does not cohere well with his preferences over the type of actions of eating juicy summer fruits. Similarly, my preference for the token action of dedicating a greater percentage of my income to my pension fund does not cohere well with my preference for actions that are non-frugal in character. It is no different than signing the lease for a timeshare in the Virgin Isles with a clever salesperson in charge – in some respect, I do want it, but in another respect, I do not, since it does not fit in with my overall preference structure. We choose on the background of a fragmented self. In answering the question whether we do or do not want to buy into the *Save More Tomorrow* scheme, whether the fox does or does not want the grapes, or whether we do or do not want the timeshare, a gloss is needed – in some respect, yes, in another respect no.

Of course coherence may be regained by making changes in my preference structure at large. The fox may well turn away from juicy summer fruits in general after

[6] This section builds on Bovens (1992).

[7] This is the preference change interpretation of the fable, which we find in Elster (1983, p. 123). This interpretation differs from a case of self-deception in which the fox would turn away and say that *these* grapes are too sour, are unripe, or what have you. This would be a case of a belief change induced by considerations of feasibility rather than by evidence for the proposition at hand.

a few bad experiences with fruit that is beyond his reach. Similarly, I may acquire a taste for frugal actions after a few *Nudges* in this direction. There are various mechanisms that may bring this about. I may come to appreciate such actions through discovering hitherto unknown attractive features. I may become habituated in Aristotelian style – my feelings may simply shift by repeatedly acting frugally or eating healthy foods. I may come to self-identify as a person who acts frugally or eats healthy foods on grounds of cognitive dissonance. All such mechanisms could be successful in bringing about preference shifts over action types and then my newly acquired preference is genuine.

What can we say about *Nudge* in cases of *Ignorance*, *Inertia*, *Akrasia*, of *Queasiness*? In these cases *Nudge* steers us in the direction of what we consider to be in line with our overall preference structure. Our initial preferences over action tokens do not cohere well with our overall preferences. So now we are *Nudged* in a direction that restores coherence between our actions and our overall preference structure. Is this a genuine preference change?

There is another lesson to be learned from the fox. Suppose that the fox tells us upon sincere introspection that he does not want the grapes anymore. Suppose that he even changes his overall preference structure and turns up his nose for juicy summer fruits in general. But now suppose that we lower the vine for the fox. Would he reconsider? Suppose that he would, maybe not immediately, but after encountering a few low-hanging bunches of grapes, he would be at it eating grapes again. Then we would need to qualify the claim that the fox changed his preferences. Again, we would add a gloss – we would say that his preference change was too short-lived to qualify as a genuine preference change. Similarly we would be hesitant to say that the *Nudge* made us prefer to dedicate a greater percentage of our income to our pension fund, when we would decide differently without the aid of some clever choice architecture. Even though we may have an overall preference to be more frugal, we can hardly be said to genuinely prefer to dedicate a greater percentage of our income towards our pension fund next year when this type of action is not resilient without the aid of *Nudges*.

As before, it may be the case that a few *Nudges* provide me with a taste for frugal actions and that it becomes easier for me to act in accordance with my overall preference structure, also in the absence of clever choice architectures. Then the non-resilience objection drops out and the preference change becomes genuine.

Let me sum up by making the point by means of *Cafeteria*. Suppose that a *Nudge* succeeds in making me take the healthy snack. Did it then induce a preference in me for the healthy snack? In some respect, yes – I revealed my preference through my choice. But this may need a gloss. Suppose that I am actually the person who values the life style of the glutton. Then in another respect, I do not genuinely prefer the healthy snack. This case maps onto the cases of *Exception* and *Social Benefits*. Or suppose that I do prefer a healthy life style, but I continue to take ice-cream under non-*Nudge* condition. Then you would look at me strangely if I were to proclaim that I genuinely prefer the healthy snack when I am placed under *Nudge* conditions. You would point out to me that I just took the healthy snack because of the choice architecture – I do not genuinely prefer the healthy snack. This case maps onto

the cases of *Ignorance*, *Inertia*, *Akrasia*, and *Queasiness*. However, if adjustments to the overall preference structure are made in *Exception* and *Social Benefits* or resilience is gained in *Ignorance*,..., *Queasiness*, then the preference change becomes a genuine preference change and we no longer need glosses.[8]

10.6 Does *Nudge* Build Moral Character?

We continue with cases of *Ignorance*, *Inertia*, *Akrasia*, and *Queasiness*. It is a lack of self-control that blocks us from acting in accordance with our overall preference structure. So when we are *Nudged* in the direction of actions that we take to be in our interest all things considered, does this build moral character? Does it increase our capacity for self-control?

The folk singer Karen Dalton once said that she sang softly because she wanted people to listen to her. This strikes us as paradoxical. Certainly people are more likely to listen when you raise your voice. Indeed, this is the expectation of the short-term effect. But the long-term effect may be precisely reversed. Think of a grade school teacher who is prone to raise her voice. This may be effective in the short term, but she may have to raise her voice more and more and the overall effect may be that more children would have listened to her had she never raised her voice to begin with. Similarly, there is research showing that the death penalty has a deterrence effect in that the rate of pardon by the governor correlates with the rate of violent crime in subsequent years (Gittings and Mocan 2003). This is consistent with the brutalisation effect – capital punishment may contribute to a more violent culture and may increase violent crime in the long run.

Now it may be the case that repeated *Nudging* in public health and pension funds may have short-term positive effects at best. *Nudging* may not create sustainable effects on people's behaviour for the long-term; as time goes on, the level of *Nudging* required to retain this effect may increase. Just as Karen Dalton did not want to raise her voice, knowing full well that some people would zone out, we should not

[8] I would like to flag the following nagging concern. Some people may object that if we regain coherence through changes to our overall preference structure (in *Exception* and *Social Benefits*) or to our particular preferences under non-*Nudge* conditions (in *Ignorance*, ..., *Queasiness*) through mechanisms such as habituation, cognitive dissonance, ..., then this is worrisome because of the broad scope of non-autonomous preference change. (This objection was raised by Jason Alexander and Alice Obrecht.) In "Sour Grapes and Character Planning" (1992), I argued in response to Jon Elster (1983, pp. 24–25) that it is not the lack of autonomy of the fox's preference change, but rather the lack of coherence between his adjusted preference and his overall preference structure. And this is what distinguishes sour grapes from character planning. In character planning, there is an adjustment of the particular preference as well as an adjustment of our overall preference structure so that coherence is restored. The lack of autonomy in preference change is unproblematic – our preferences may be fully rational even though we did not autonomously acquire them. I am comfortable repeating this line when it comes to *Ignorance*,..., *Queasiness*, but slightly nervous when it comes to *Exception* and *Social Benefits*. But elucidating this difference – if indeed there is such a difference – will require more reflection.

be lured by the short-term success of *Nudging* either. To warrant long-term success, we should let people make their own decisions while providing minimal aid. My point is that short-term success of *Nudge* may be consistent with long-term failure. The long-term effect of *Nudge* may be infantilisation, i.e. decreased responsibility in matters regarding one's own welfare.

But of course things ain't necessarily so. Cognitive dissonance, habituation, acquiring a taste for the good-making features in *Nudged* actions may bring about long-term preference change as well. More people may come to adjust their overall dietary habits (and not only in the *Cafeteria* setting) or become more prudent in general (and not only in the *Save More Tomorrow* scheme.) This brings us back to the question of whether *Nudge* induces genuine preference change. When we come to acquire a taste for the *Nudged* actions, then the effects will be more broad-ranging and long-lasting.

At the end of the day, different people will be affected in different ways and it is an empirical question whether there does exist something like the infantilisation effect, just like it is an empirical question whether there exist something like the brutalisation effect of capital punishment. My only aim here is to point out that, just as a study of the (short-term) deterrence effects of capital punishment by means of time-series analysis is not the last word, a study of the (short-term) success of a particular *Nudge* is not the last word either. Granted, brutalisation and infantilisation effects are difficult to study through empirical testing. It does not suffice to do cross-population studies and to point to the correlation between capital punishment and the number of executions on the one hand and the rate of violent crime on the other hand, since the causal direction is unclear. Counter to the brutalisation effect, it may well be the case that high rate of violent crimes is the cause of the institution and the prevalence of capital punishment, rather than vice versa. The same problem would occur if we were to find a correlation between some measure of responsibility and paternalistic policies. Counter to the infantilisation effect, it may well be the case that the low measure of responsibility is the cause of the institution and the prevalence of paternalistic policies rather than vice versa.

10.7 Who Is *Nudging*?

It matters a great deal who is doing the *Nudging*. Let us start with a case in which I set up a *Nudge* to constrain my own behaviour. This is an example of *sophisticated choice* (McClennen 1990, p. 12). I may force myself to decide on increased pension-fund contributions earlier rather than later because I know that it is my only hope to commit a reasonable amount. I don't think that there is much to object here. Now some strong-willed people consider it to be wrong for me to decide, say, not to bring any liquor in the home. They seem to believe that we should educate ourselves so as to become *resolute choosers*, who are able to commit just as much to their pension funds after receiving a raise as before receiving it and who are able to drink just as little, whether there is liquor in the home or not. But let us bracket such ideals of

perfectionism. As long as we are self-legislating, there seems to be little to object to in engaging a *Nudge*.

But now let us go one step further. Suppose that I choose a nudging partner and consciously or unconsciously take his or her nudging to be a good-making feature. Or suppose that I choose to work in a self-professed paternalistic company. In either case, it seems that I have little to object to when the fridge or the line with food items is carefully arranged so that I am more likely to take the healthy options.

But this brings us to the actual concern with *Nudge* as a social policy instrument. What if a majority elects a government with a nanny-state platform? Do they have a democratic mandate to *Nudge*? What about the minority who does not want the government to interfere with their preference formation? We will return to this question below.

10.8 Transparency

Without going into the empirical details, let us suppose that the use of subliminal images could actually bring about preference change. Now typically the use of such devices makes people extremely nervous. Suppose that the government starts a public health campaign to reduce obesity. Let there be a social group with problematic dietary habits. Research shows that there is a high density of viewers from this social group for a particular TV programme. So we decide to splice pictures of happy carrot-eaters into this programme as subliminal images. T&S object to this practice because of the lack of transparency (2008, pp. 244–245). But then suppose that the government simply announces that it will combat social problems by means of subliminal images. T&S object that also this would not suffice, because 'manipulation of this kind is objectionable precisely because it is invisible and thus impossible to monitor' (2008, p. 246).

So how is this any different from *Cafeteria*? We need to make a distinction between *type interference transparency* and *token interference transparency*. It is one thing for the government to say that they will be using certain types of psychological mechanisms to solve social problems. This is type interference transparency – the government is transparent about how it will try to interfere with our agency. But then there is no difference between *Nudges* and subliminal images – the government can announce that it will *Nudge* and that it will use subliminal images. Yet T&S support the former and object to the latter. So type interference transparency is not enough.

I take it that T&S also wish to have token interference transparency. How does *Nudge* differ from subliminal images? Being exposed to a particular image at a particular time is a token interference by means of a subliminal image. When we are affected by such a token interference, there is no way that we could notice (blocking the use of special equipment). But if we are being *Nudged*, it is possible to recognise here and now that the food is arranged in a particular manner, that the pension savings forms are sent early to facilitate saving etc. So in a *Nudge*, it is possible to recognise each token interference.

So does this mean we need to put up a billboard next to the food line stating: "Research shows that people are more prone to take food items displayed earlier rather than further down the line. Many of our customers are trying to lose weight but find it difficult to do so. To help them, we have arranged the snacks in the food line with healthier items displayed earlier so that they are more likely to choose these items." The problem is that these techniques do work best in the dark. So the more *actual* token interference transparency we demand, the less effective these techniques are. But it may just be sufficient that there is *in principle* token interference transparency. A watchful person would be able to identify the intention of the choice architecture and she could blow the whistle if she judges that the government is overstepping its mandate. This in principle token interference transparency is not possible for subliminal images. In giving the government a mandate to use subliminal images we would be signing a blank check and could only hope that they will not be abusing their power and splice in ads for the incumbent in the next election.

In summary, subliminal images are deemed impermissible because they do not satisfy *in principle token interference transparency*, whereas T&S-style *Nudges* do pass this requirement. But then are we not confident that there are some watchdogs with sophisticated equipment keeping an eye on the government? Certainly, but I think that we find it important that also *we ourselves* could decide to become watchful and unmask any manipulation. In the democratic process we may give the government a mandate to engage in certain types of *Nudges*. But then we wish to respect the right of minorities who do not appreciate this type of manipulation. To safeguard their interests, we stipulate that every *Nudge* should be such that it is in principle possible for everyone who is watchful to unmask the manipulation.

10.9 The Moral Permissibility of *Nudge*

I have pointed to a number of issues that are relevant when we judge the permissibility of a particular *Nudge*.

First, it is less worrisome when the *Nudge* brings our agency in line with our overall preferences, as in *Ignorance, Inertia, Akrasia and Queasiness* than when the projected agency is not in line with our overall preferences, as in *Exception* and *Social Benefits*.

Second, it is less desirable when a *Nudge* is local and leaves us with a fragmented self. We become incomprehensible to ourselves – why did we not act in line with our overall preferences or why is this kind of agency not resilient under non-*Nudge* conditions? We can avoid such a fragmented self if our *Nudging* brings about change in our general preference structure or change in our agency that continues to hold under non-*Nudge* conditions. But there is a tension here. Some will undoubtedly be even more worried if *Nudge* brings about massive changes in our preferences through psychological mechanisms such as habituation, cognitive dissonance etc. Fragmentation avoidance comes at the cost of even more non-autonomous preference change. This may be worrisome in cases like *Exception* and *Social Benefits* (see footnote 8).

Third, *Nudge* is less desirable when it creates a people who have become incapable of taking their lives in their own hands and to make autonomous changes in their agency to make it fit in with their overall preference structure. Such long-term infantilisation effects are difficult to assess empirically but it is nonetheless a concern that does not go away. Adam Smith (Part VI, Section III; pp. 143–145) thought that adversity was the best school to develop the respectable virtue of self-command. The cost of *Nudge* may be that we forego the chance to gain the virtue of self-command.

Fourth, the less control we retain over being *Nudged*, the more problematic it is. If we choose to put ourselves into a situation that is rich with *Nudges*, then we have little to complain about. But does this type of consent extend to a democratic mandate to the government to be *Nudged*? I have argued that *Nudges* must be transparent in principle at the level of each token *Nudge*, in order to ensure that everyone can unmask the manipulation if they wish to do so. This protects the rights of the minorities who do not wish to be so manipulated and it keeps a check on the government.

There are many other factors that enter into the permissibility of *Nudge*. Let me just flag a few. Advertisement for products that do not increase welfare may use all kinds of *Nudge* style techniques and the government may be fighting a losing battle against, say, obesity, if it cannot access the same arsenal of techniques. Furthermore, governments commonly set up quasi-markets to increase efficiency in the provision of public goods. Citizens are bombarded by technical information from competing providers. Securing health insurance should not be as complicated as choosing a cell-phone. If the government institutes such quasi-markets then it also has the responsibility to navigate people through them which may involve more or less gentle *Nudging*. Finally, the more urgent the problem that a *Nudge* is trying to tackle, the less it meets with qualms. Instituting *Save for Tomorrow* may be more acceptable in the US than in South East Asia, considering differential saving rates. Instituting *Cafeteria* may be more acceptable in Chicago than in Paris, considering differential obesity rates. And, no doubt, in assessing the permissibility of *particular Nudges*, many more considerations that are idiosyncratic to the case at hand will emerge and each case will need to be assessed on its own merits.

References

Bovens, L. 1992. Sour Grapes and Character Planning. *Journal of Philosophy* 84: 57–78.
Bovens, L. 1995. The Intentional Acquisition of Mental States. *Philosophy and Phenomenological Research* 55: 821–840.
Chakrabortty, A. 2008. From Obama to Cameron, Why Do so Many Politicians Want a Piece of Richard Thaler? *Guardian* July 12.
Elster, J. 1983. *Sour Grapes*. Cambridge: Cambridge University Press.
Gittings, R. K. and N. Mocan. 2003. Getting Off Death Row: Commuted Sentences and the Deterrent Effect of Capital Punishment. *Journal of Law and Economics* 46: 453–478.
Hume, D. 1978. *A Treatise of Human Nature*. (2nd Ed.) Edited by L.A. Selby-Bigge. Oxford: Clarendon.

McClennen, E. F. 1990 *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.
Savage, J. 1954. *Foundations of Statistics*. New York: Wiley.
Smith, A. 1968. *The Essential Adam Smith*. Edited by R. L. Heilbroner. New York: Norton.
Sunstein, C. R. and R. H. Thaler. 2003. Libertarian Paternalism Is Not an Oxymoron. *University of Chicago Law Review* 70: 1159–1202.
Thaler, R. H. and C. R. Sunstein. 2003. Libertarian Paternalism. *American Economic Review* 93: 175–179.
Thaler, R. H. and C. R. Sunstein. 2008. *Nudge*. London: Yale University Press.