Rethinking Nudges

by Philippe Mongin and Mikaël Cozic
Centre National de la Recherche Scientifique and HEC Paris.
Département de philosophie, Université de Créteil, and Institut Universitaire de France.

May 2014 (provisional version)

1. Introduction

In the recent discussions of economics and philosophy, nudge refers to a policy for redirecting an agent's choices by slightly altering his choice conditions, in such a way that the interference with these conditions is minimized. In their now well-known eponymous book, Thaler and Sunstein (2008) introduced nudge as a separate category of interventions by either public or private parties, illustrating its meaning by way of examples rather than analyzing it theoretically. They also made a case for benevolent nudging, i.e., nudging intended to promote the agent's welfare, as a substitute for, or complement of, benevolent interventions that rely on bans, commands, or heavy manipulations of choice incentives. Moreover, they claim that benevolent nudging reconciles liberalism (in the sense of essentially respecting the individual’s freedom of choice) and paternalism (in the sense of giving priority to the welfare improvement over the individual's spontaneous will). Hence the brand name - libertarian paternalism - of the moral doctrine they promote, both in the well-read book and in earlier, more scholarly pieces.¹

The nudge concept is also currently debated at the fringes of behavioural economics, which suggests a different connection of ideas. Thaler and Sunstein are again the source. Both in their book and elsewhere, they have emphasized the agents' limitations of rationality, borrowing extensive samples from Kahneman and Tversky's work among others, and they

¹ Nudge (2008) is usefully supplemented by more academic work; see especially Sunstein and Thaler (2003) and Thaler and Sunstein (2003). References below are to Nudge (2009) - the paperback slightly expanded edition.
have also defined nudges as being interventions that make strategic use of these limitations. This is not all there is to nudges and bounded rationality. The latter could very well enter the former by relating to the target of the intervention, whether it also relates to the chosen means or not. This possibility concerns a benevolent party who is worried about the unfavourable consequences of bounded rationality for the agent. It could try to counter them in whatever way, but *Nudge* singles out the interesting particular case in which it would also instrumentally resort to bounded rationality, thus turning the latter against itself as it were. ²

In sum, nudge has three distinctive senses, (1) as an intervention that interferes with the choices conditions minimally, (2) as an intervention that uses bounded rationality strategically, and only in the case of benevolent nudging, (3) as an intervention that tries to remove the negative effects of bounded rationality. We will call them *nudge 1*, *nudge 2* and *nudge 3* for brevity. That nudge means more than one thing matters a great deal to Thaler and Sunstein's defence of liberal paternalism. They need to show that benevolent nudges effectively improve the agents' welfare despite being such minor interferences, and what they argue in effect is that behavioural economics makes exactly for the right sort of interventions. They presume that it is less intrusive and no less efficient to make clever use of the agents' bounded rationality, especially when turning it against itself, than it would be to devise incentives according to a standard rationality pattern. This is how they hope to remain liberal, while they endorse paternalism, in the sense of overruling the agents' spontaneous choices for their own good.

Among Thaler and Sunstein's preferred examples of benevolent nudges are changing the display of food in public canteens, proposing plans for arranging one's future savings, allowing for withdrawal periods in consumers' choices, translating complex data into pragmatically usable information, enlarging the choice among complicated items with suitable default options. From what they say or suggest, these interventions aim at circumventing the obstacles to a good decision that cognitive biases and related features involve, exploit these very weaknesses for their success, and are so well devised that they do

² Within behavioural economics itself, there is a tendency to base policy recommendations on its findings. See in particular the plea for "asymmetric paternalism" in Camerer, Issacharoff, Loewenstein, O'Donoghue and Rabin (2003). A comparison with Thaler and Sunstein's work would be in order but goes beyond the scope of this paper.
not meddle severely with the initial decision processes. That is, each example in the list supposedly satisfies all senses of nudge together – nudge 3, 2 and 1 in that order.

In the budding literature that Thaler and Sunstein have inspired, the normative assessment occupies centre-stage. Some have evaluated nudges per se, i.e., independently of their alleged connection with liberal paternalism, but most discussants have rather been concerned with scrutinizing - and indeed generally rebuking - the possibility of this doctrine. The attention that writers in either group pay to the multifarious sense of nudge varies, but even those who noticed it did not explore it in detail. This is what we propose to do here. We are primarily concerned with a better understanding of the new concept, and it is only after having tried to separate its constituent ideas that we will return to the normative questions and decide whether or not Thaler and Sunstein made a plausible case for libertarian paternalism.

The next section is a semantic analysis of the nudge 1, 2 and 3. There is no connection between them at the level of meaning alone, so that Thaler and Sunstein's enterprise depends on a factual claim, i.e., that there exist interventions satisfying the three concepts - or at least the first two, because this would already provide a sufficient basis for their libertarian paternalist claim. As section 3 shows in detail, Thaler and Sunstein's own list of examples defeats this factual claim, whence the presumption that it is false. This is the core argument of the paper. Section 4 relies on it to reconsider the issues of liberal paternalism.

2. Nudge 1, 2 and 3

Compare the following two sentences from Thaler and Sunstein:
"A nudge, as will use the term, is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any option or significantly changing their economic incentives" (2009, p. 6).
"In accordance with our definition, a nudge is any factor that significantly alters the behavior of Humans although it would be ignored by Econ" (2009, p. 8).

---

3 See especially Bovens (2009).
4 Hausman and Welch (2010), and Selinger and Whyte (2011) after them, have a clear sense of the full range of meaning of nudge. It does not come out so transparently from Mitchell (2005), Grüne-Yanoff (2012) and Qizilbash (2012).
The two sentences correspond to the nudge 1 and nudge 2 senses, respectively. When read together, these two sentences appear to be puzzling, because each claims to state a definition, and there is no way in which they can be made equivalent analytically, i.e., by resorting only to the meaning of the words and the laws of logic.\(^5\) In the elementary view of definitions, the *definiens* consists of necessary and sufficient conditions for the *definiendum*, so to define something in two ways commits one to prove a logical equivalence between them. On a more liberal view, a definition can be partial, with the *definiens* providing only a necessary or a sufficient condition for the *definiendum*, which would not make illogical to have nonequivalent definitions, but the methodological practice firmly excludes this. Thus, however definitions are logically conceived of, any condition that is not equivalent to the *definiens* can be imposed on the *definiendum* only through a factual claim, not by another definition.

This last comment provides the way out of the little semantic conundrum into which Thaler and Sunstein have trapped themselves. We could say that they take one condition to be the *definiens*, and that they make the factual claim that nudges so defined typically satisfy the other condition. If one takes the theory to have a primarily normative aim, it seems appropriate to take nudge 1 as being the definition, since it is closer to the normative than nudge 2. It describes a kind of interventions that libertarian paternalism would approve of, thus providing an intermediate-level criterion to evaluate interventions. Nudge 2 would describe factual properties that are instrumentally useful to meet the criterion. As to the opposite solution, which takes nudge 2 to be the definition, it seems to put the wrong emphasis.\(^6\) However, a symmetric and non-committal semantics is sufficient for the purposes of this paper. We will eschew the definition problem, just talk of *interventions*, assuming that this has a sufficiently clear meaning, and interpret Thaler and Sunstein as making the factual claim that there is a relevant range of interventions that jointly satisfy the nudge 1 and nudge 2 properties. Nudge 3 enters the discussion mostly as a complication, being concerned with only *benevolent* interventions and a highly specific class of them for that matter.

Before checking Thaler and Sunstein's factual claim, we need to analyze, and if necessarily revise, the very properties of nudge 1 and nudge 2. Regarding the former, from what the

\(^5\) We gloss over the classic Quinean difficulties of this analyticity concept.

\(^6\) The nudge concept entered stage late in the authors' joint work, which has been concerned with libertarian paternalism all along.
authors write, they are (i) not forbidding any option and (ii) not significantly changing the economic incentives. Both are intended to serve as proxies for freedom of choice, a major concept for liberal paternalism, which we may ignore now as we do this doctrine. We introduce some decision-theoretic ideas to make coherent sense of (i) and (ii). We will assume that the agent faces a currently given set of choice objects, each of which he identifies with a vector of qualitative or quantitative values for relevant attributes, these being taken from a currently given list. For example, he is faced with a set of national lottery tickets that he identifies with three-dimensional vectors of prices, possible gains, and probabilities of gain. To pick out one or several objects from the choice set, the agent compares their attribute values according to various criteria, which also taken from a currently given list. We should be non-committal here, only assuming that, for each attribute, at least one criterion applies; otherwise the attribute would not be taken in consideration. Everything else will depend on the agent and the case: there may be more than one criterion applying to an attribute, or more than one attribute being in the scope of a criterion. When a national lottery ticket is evaluated in terms of its net expected gain, a single holistic criterion applies to the object; when it is evaluated in terms of its net expected subjective value, a criterion first applies to the money components, and then a holistic criterion applies to the object thus transformed; and it is also possible that each of the three dimensions has a distinctive criterion and a non-standard holistic criterion applies, as some experimental evidence indirectly suggests. 7 Neither the criteria individually, nor the holistic criteria that aggregate them need be orderings. With this sketch, we do not account for the search part of the decision process, since we suppose a given choice set, a given list of attributes, and a given list of criteria, but we still allow for some significant features of bounded rationality. 8 By and large, they are concerned with the preference side of the decision process. The structure imposed by the attributes and criteria may be so weak that the resulting preferences over choice objects are incomplete, and it may be so strong that these preferences are inconsistent. 9

---

7 The celebrated experiments on preference reversals can be interpreted in this way; see Lichtenstein and Slovic (1971), Tversky, Kahneman and Slovic (1990) and many others.
8 Simon (1982), who introduced the bounded rationality concept, was also the first to distinguish between the search problem of choice and the problem of exploiting the determinants of choice appropriately.
9 It is a subtle question, not to be discussed here, whether incomplete and inconsistent preferences still count as genuine preferences. See in particular Hausman (2012).
This sketch should be supplemented with a methodological precept. When the choice set is modified, as a result of an intervention or for any other reason, the outside observer of the new choices should try not to upset the assumptions on what counts as an attribute, and on how the criteria work on them. If possible, observed changes in choices should be traced to the change in the external conditions of choice, taking the internal conditions to be fixed. Arguably, this precept underlies the empirical work inspired by any sustained conception of choice, both in standard and behavioral economics. Putative regularities may connect changes in the internal conditions with changes in the external conditions, and changes in observable choices with changes in the internal conditions, but they are more difficult to uncover than those connecting observable choices with changes in the external conditions, since the last ones can be experimented with more easily than the others. The principle is so trite as to go unnoticed by today's behavioural economists even if their practice endorses it.

With these preliminaries at hand, we see that even the simple idea of not forbidding any option is equivocal. For sure, it means that the choice set either remains the same or increases as a result of the intervention, but in the latter case, the added objects may or may not be similar to the earlier ones, and this can make a significant difference. If they are similar, the methodological principle applies sensibly, but if they are not, one may have to concede that new attributes or new criteria enter the decision process along with the new objects, which amounts to giving it up.

Now to the not so simple idea of not significantly changing the economic incentives. The incentive concept has no specialized location in the sketch. Prima facie, it can relate to all of its component parts, i.e., the criteria, the attributes and the choice set itself. When the latter is modified, the methodological principle suggests that this change represents the change in incentives, everything else remaining the same, but this becomes implausible with, say, education and advertisement. No doubt, these familiar interventions are best construed as changing the incentives through changing attributes and criteria, and not simply choice sets.

If we read Thaler and Sunstein correctly, they insist on (i) and (ii) being met together: "without forbidding any option or significantly changing (the) economic incentives" means "neither forbidding any option, nor significantly changing the economic incentives", and not, as a less exacting concept of nudge 1 would require, "not forbidding any option, or not changing their economic incentives". At the same time, Thaler and Sunstein also insist on
"(altering) people's behavior in a predictable way". Only by uncovering solid facts in choice behaviour can one make it predictable, and the common way of discovering them is to apply the methodological principle, as behavioural economists themselves do. But systematic application of the principle make the two conditions potentially conflicting. An example will bring this critical point home.

Take the standard, neoclassical principal-agent theory of the insurance market. It consistently applies the methodological principle by analyzing changes in the agent's incentives in terms of changes in the choice set; what counts as the internal conditions here, i.e., the agent's well-structured preferences, remains fixed by assumption. Thus, the way an insurance company is supposed to counteract moral hazard is by either ordering more ex post inspections or by ex ante altering its menu of insurance contracts. The latter move will typically involve suppressing older contracts, and not simply adding new ones. This is still true if, for some reason, the company aims at producing only a small change in the agent's incentives. Suppose that it has thus far issued natural flood insurance contracts, both with and without deductibles, and having realized that customers did not make prevention efforts, it wonders on how to incite them to that. The solution, which is to offer all contracts with deductibles, is the same whether the planned incentive change is large or small. This difference only affects the size of the deductible, and in any case, the initial menu should be curtailed of some of its contracts.

This little example teaches that the two characteristic properties (i) and (ii) may clash with each other. Of course, this possibility depends on how these two properties are rendered in the decision-theoretic framework, but a ceteris paribus assumption, such as is made in the neoclassical analysis of the example, fits with Thaler and Sunstein's predictability requisite. As a consequence, the nudge 1 concept appears to be overdetermined.

The nudge 2 concept also suggests second thoughts. Thaler and Sunstein express it by contrasting "Humans" with "Econs". This reflects their adherence to the programme of behavioural economics, which is to replace the rationality assumptions of standard decision theory and neoclassical economics by descriptively more realistic assumptions, and ideally, by experimentally based assumptions. The general public having itself long entertained doubts on the "rational economic man", Thaler and Sunstein's formulation is a walkover to rhetorical success. However, as the record shows, standard theories have the resources to account for some of the rationality failures brought to light by experiments, and the
alternative assumptions that behavioural economics has floated are not powerful enough to
displace the rationality assumptions entirely. The specialized work in today's theory of risky
decisions exemplifies either point well. ¹⁰ In the present state of the art, "Humans" and
"Econs" are overlapping populations, and to contrast them with each other is not the proper
way of introducing nudge 2.

Moreover, is it is the agents' characteristics, not the clash between different schools of
thinking about them, which is at stake, and even if standard theories were clearly hopeless,
and behavioural economics were clearly promising, there would be little reason to stress this
epistemological contrast. Let us then translate nudge 2 into the language of objective
properties. ¹¹ The literal translation is that the intervention succeeds only with boundedly
rational agents, not with perfectly rational ones. This is still too theoretical a distinction, since
real agents are not so segregated, and what matters in actual fact are not the individuals, but
the decision processes they follow under given circumstances. Hence we suggest, nonliterally,
that an intervention counts as nudge 2 if it strategically exploits bounded rationality within
the decision process. This can take the form of either bringing about a new process or
modifying an existing one, but in the latter case, not to the point of inducing perfect
rationality. Otherwise, nudge 2 interventions would cover strategies that are not reproducible,
since they destroy the bounded rationality basis of their efficacy on the first use.

How should one conceive of bounded rationality in the present context? The decision-
thetoric sketch is not so useful here, because it takes for granted the search part of the
decision process, and this should be included, and perhaps even emphasized, when it comes to
nudge 2. It also has the defect of being purely static. For want of a more sophisticated
strategy, we take up the list of psychological factors that Nudge relates to bounded rationality
and criticize it. The list has three broad groups: "biases and blunders", "resisting temptation",
and "following the herd", which we briefly consider in turn.

¹⁰ For example, today's well-regarded cumulative prospect theory is an admixture of relatively
standard assumptions with some experimental economics findings, and it has become
common ground between the two camps. See Wakker (2010) for a recent account.
¹¹ Admittedly, Thaler and Sunstein do not stop at "Humans" and "Econs" and also mention
the contrast of today's cognitive sciences between "system 1" and "system 2". See Kahneman's (2011) introduction to this popular theme. We do not think it is really usable
here.
Thaler and Sunstein are aware that the third group may be out of scope. Apparently gregarious behaviour can often be justified in terms of strategic equilibria, which may even be based on explicit strategic thinking on the agents' part. Thaler and Sunstein do not express much doubt about the second group, although they should have. Standard decision theory allows for ex post revisions of ex ante plans when new information occurs, and some of the behaviour that is informally described as not resisting temptation is of this revision type. The theory has also considered the possibility that information remains the same, but even in this case, time-consistency does not have a clear rationality status. There is an ongoing debate on this condition in the context of expected utility theory, with some scholars claiming that it is a rationality condition and others disagreeing. Furthermore, the issue of time-consistency must be kept distinct from discounting issues, since the latter involve a psychological content that rationality does not have to consider. Thaler and Sunstein's informal analysis of the temptation problem occasionally jumps over this basic distinction.

By contrast, the first group is indisputably relevant to bounded rationality, with its classic list of biases: anchoring and adjustment, availability, representativeness, overconfidence, loss aversion, status quo bias, framing. With some effort, the sketchy decision-theoretic model could accommodate them by including overconfidence, loss aversion and status quo bias into the criteria, and by including anchoring and adjustment, as well as availability and representativeness, into the attributes. But the model is powerless on framing, which suggests that this item stands alone in the list. Another internal distinction is that some factors can be viewed as heuristics as well as biases, while others appear to be straightforward biases. Tversky and Kahneman (1974) introduced anchoring and adjustment, availability, and representativeness in the context of judgment under uncertainty, and argued that they were common heuristics to replace or simplify probability calculations. They also described them as biases since they lead to errors in many cases. By contrast, there is no heuristic sense to be made of overconfidence, loss aversion, status quo bias. Framing is an outlier again since it is not a heuristics and perhaps not even a bias; indeed, a bias goes always in the same direction, whereas framing errors have no general pattern.

---

13 Machina (1989) surveys the issues. A major problem is that in the presence other conditions, which may seem cogent, time-consistency entails the expected utility criterion, which some, like Allais, have complained might lead to violations of rationality.
The secondary literature on nudges has expressed concerns about the reliability of bounded rationality phenomena for the purpose of an intervention. The internal distinctions just drawn suggest that these concerns should be hierarchized. Biases that result from heuristics seem to be less reliable in their workings than straightforward biases. The reason is that, unlike the latter, the former are optional and they are not fully unconscious; indeed, they may even be part of an explicit deliberation. They can be given up, whereas it is difficult to correct one's pure biases. Framing is at the bottom of the reliability ladder.

In sum, nudge 2 goes in the opposite direction to nudge 1, i.e., underdetermination. Its conceptual meaning is vague, and its empirical extension is unclear beyond the received list of biases. This situation is easily explained by the fact that today's behavioural economics offers more hints than genuine results. It has at least the advantage of leaving some hope that nudge 2 and nudge 1 interventions will overlap.

We follow Thaler and Sunstein once closely by considering interventions more generally than benevolent ones. Benevolence here means that the intervention aims at increasing the agent's welfare, as represented by the acting party; welfare may not be the sole subjective aim, but it must be more than a side-effect; and it must bear some relationship to the agent's actual sense of welfare, even if discrepancies are a common motive for intervention. There are so many problems surrounding these ideas that it seems best to discuss them separately from nudging interventions as generally conceived. Thus, for the latter, we will only suppose that the intervening party has an interest in reorienting the agent's choices. This may be a benevolent interest, but possibly also a mischievous or indifferent one. In this way, we do some justice to Thaler and Sunstein's recurring claim that agents are often nudged by others without any consideration for their good, and quite possibly at their expense.

Our last technical concept, nudge 3, enters stage at this juncture. It covers those benevolent interventions which proceed by countering the unfavourable effects of the agent's bounded rationality on his decisions. There are other ways in which the objective of improving the agent's welfare could be reached, even if the agent has limited rationality, and even if the intervention also consists in redirecting his choices, so this is but a very particular case of

---

15 See, e.g., Hausman and Welch (2009) and Grüne-Yanoff (2012).
16 These problems are central to Qizilbash's (2012) critique.
benevolent interventions. Nudge 3 should carefully be distinguished from nudge 2, the latter being concerned with bounded rationality as a means, while the other includes it into the objective (though of course negatively, as an obstacle to welfare). However, the distinct concepts are related in the following way. If one argues against attributing nudge 2 to an intervention by saying that the decision process is free from any kind of bounded rationality, one also has a reason against attributing nudge 3, very simply because bounded rationality has no role at all to play in the analysis.17

Nudge involves some secondary equivocations that are easy to dispel. First, by the same word, the authors refer both to interventions and the biases and decision faults underlying some of them, as in nudge 2. The glide is explainable, but it is best to avoid it, as there is so much more to the idea of a mechanism used with a purpose than there is to the idea of this mechanism taken in and by itself.18

Second, Thaler and Sunstein think of nudging sometimes as being successful (in the sense of altering behaviour according to the purpose) and sometimes as being only intentional (it may either fail or succeed). This double meaning is shared by many words of the intentional language,19 but it is barely acceptable when an overall assessment of the intentional action is needed. What is at issue with nudging is not only whether it is morally acceptable, but also whether it is instrumentally efficacious, and there is more than one way it could fail in this respect. This whole discussion should not be precluded, and we will thus conceive of interventions in the wider of the two senses permitted by the intentional language.20

3. Categorizing interventions

17 Our classification of nudges complicates the available pictures. We read Bovens (2009) as defining nudge only like our nudge 2. By contrast, Hausman and Welch (2010, p. 126) define nudge by the conjunction of nudge 1, 2 and 3: "To sum up: Nudges are ways of influencing choice without limiting the choice set or making alternatives appreciably more costly in terms of time, trouble, social sanctions, and so forth. They are called for because of flaws in individual decision-making, and they work by making use of these flaws". The last sentence gives bounded rationality a dual role in the means and the objective.
18 This complaint has already been made in the literature.
19 See Ryle's (1949) discussion of "success words".
20 Commenting on paternalism, Hausman and Welch make a related comment: "What characterizes paternalism are the aims with which one acts and the means one employs, not whether one is successful" (2010, p. 129).
This section confronts Thaler and Sunstein's main examples of nudges with the semantic requirements put on nudge 1, 2 and 3. The guiding question will be whether any of these examples fits the corresponding sets of requirements together. The answer will be facilitated by the previous observation that the nudge 2 and 3 properties fall together when no bounded rationality feature is involved.

Let us start with the didactic introductory example of *Nudge* – the by now famous cafeteria example. The head of a cafeteria decides to improve consumers' health by rearranging the display of foods in such a way that they will be encouraged to take relatively more of the healthier ones; these foods will be made, say, more conspicuous and easier to take away. Thaler and Sunstein do not cite any experimental evidence that straightforward changes of display have any effect under related circumstances, and the example must accordingly be judged on its intuitive plausibility, which depends on how it is made precise. A changed consumption is not unlikely to take place if the customers did not previously notice the healthier food, but then the intervention is best analyzed as a rational adjustment to the new choice set, and this contradicts nudge 2. It is much less likely that consumers who previously knew about the healthier foods will modify their consumption, but let us assume for the sake of the argument that some do. In terms of our model, this can be explained by supposing that a new attribute becomes relevant (i.e., what food the cafeteria tends to recommend) or by supposing that an existing attribute is evaluated differently by an existing criterion (i.e., the ease of access to the food). The incentive change captured in this way can be said to be small, and the set of options is physically the same; so both nudge 1 conditions are met. But it is unclear whether the intervention should count as nudge 2. It may be rational to follow what is perceived to be a suggestion from a benevolent organization, and it may be rational to react to a weaker request of effort, even if the change is light. A marginalist "econ", who is always on the alert, is more likely than an inert "human" to respond to these very slight signals. If the example seems to lack plausibility, this is perhaps - an odd suggestion - because it fits with the neoclassical stereotypes better than with the behavioural ones. Now, the reasons for doubting nudge 2 also make nudge 3 dubious since we claim in essence that the example is more plausible with completely rational customers. The bounded

---

21 Marketing evidence is not directly applicable, because it usually considers the effect of changing price as well as the display (as in "loss leading" strategies).

22 Even nudge 1 is questionable, for the newly perceived choice element may bring about a significant incentive change.
rationality that could be considered at best involves the borderline phenomenon of time inconsistency and not the core list of biases.

Are the more realistic examples of _Nudge_ any more convincing? First, consider Save More Tomorrow (SMT). This is an ingenious saving scheme that commits the participants to increase their savings at future dates, with the added contributions timed to coincide with the participants' later raises in earnings. To give maximum chance to Thaler and Sunstein, we assume that the agent is already willing to save in the future exactly as SMT procures. How would one analyze the effect of making SMT available to such an agent? On the presumption that none other leaves the choice set, it makes a net addition. Incentives are not upset, as we the saving pattern is already judged desirable and the role of SMT is only to make it easier to implement. Hence the intervention qualifies as nudge 1, but of course only _ex ante_, since the agent who adopts SMT will be faced with a reduced choice set when his future earnings materialize.

We must also keep an eye on the timing when we evaluate whether the intervention satisfies nudge 2. _Ex ante_, our supposed agent will not adopt SMT if he is not worried by his possible time inconsistency, and this requires a sophisticated, and arguably a highly rational, attitude.\(^ {23} \) This contradicts any nudge 2 attribution. Claiming that time inconsistency belongs to bounded rationality does not help simply because - regardless of whether or not this is theoretically justifiable - time inconsistency is not the psychological mechanism that makes the intervention effective.\(^ {24} \) Evidently, time inconsistency must not be confused with the preventive attitude towards it; the two concepts differ even extensionally since not every time-inconsistent agent finds it worrying. Thus, when viewed _ex ante_, the intervention is nudge 1 but neither nudge 2 nor nudge 3. When the intervention is viewed _ex post_, we know

---

\(^ {23} \) Since the time of Strotz (1955-1956), the decision-theoretic literature has discussed self-commitment strategies, often leaning towards the conclusion that they are prompted by rationality. McClennen (1990), and more recently Verbeek (2007), review some of the issues. Philosophers often discuss them under the label of _weakness of the will_ (after Davidson, 1980).

\(^ {24} \) The same comment applies to inertia and loss aversion, which Thaler and Sunstein (2009, p. 114) also propose as reasons for one's committing to a SMT plan. Actually, a completely inert chooser would stick to one of the earlier plans, and perhaps also a loss averse one if SMT happens to be more costly.
that it does not satisfy nudge 1; thus even though it plausibly satisfies nudge 2 and 3, all properties are never met jointly. 25

Second, take the imposition of a withdrawal or "cooling-off" period to some consumers or borrowers. Many countries have made this mandatory in cases where the agent may be influenced (as in door-to-door sales) or is about to commit large amounts to money (as with mortgage loans). The obvious rationale is to avoid impulsive and unreflective decisions. This is bounded rationality all right, but the intervention makes bounded rationality the object of the intervention, not its means, i.e., it is nudge 3 and not nudge 2. (This is one of several cases in which Thaler and Sunstein have fallen into an equivocation concerning the role of bounded rationality.) The intervention reduces the agent's choice set dramatically, so it is not nudge 1 either. It would not do saying that the agent's decision is only postponed; this is loose wording because possible decisions differ according to the time, and the current decision to buy is clearly blocked. So we have turned another of Thaler and Sunstein's examples into a counterexample, and regrettably of the worst possible kind, because it satisfies neither of the two important senses of nudge. Interestingly, the authors seem to have been reluctant to include withdrawal periods in their list, recognizing that they include a ban and that this contradicts nudging. 26

Third, consider Thaler and Sunstein's recurring advice of implementing RECAP (Record, Evaluate, Compare Alternative Prices) into commercial law. Agents taking out insurance, contracting for mortgage loans or even making basic utility subscriptions are likely to be faced with too complex information for them to process it entirely well. What the authors call RECAP is a regulation of disclosure practices; it requires sellers to translate the technical and financial data of their products into information that is pragmatically relevant to consumers. Physical units will be turned into money figures, one-shot payments will be separated from regular ones, renewal clauses and exit penalties will be stated transparently, the delays to get the benefits or other time constraints will be estimated, and so on. To give legal force to RECAP is to impose bans, but on producers, not on consumers, who are the only concern

25 This discussion is sufficient as far as the analysis of SMT in Nudge is concerned. However, Benartzi and Thaler (2007) and Benartzi, Peleg and Thaler (2012) have a more elaborate treatment, which should also be considered.

26 See the inconclusive discussion in Thaler and Sunstein (2009, p. 253-254). By contrast, Sunstein and Thaler (2003, p. 1186-89) mentioned mandatory cooling-off periods as if they were an unproblematic example of libertarian paternalism.
here. Does this intervention satisfy nudge 1? Since the choice set remains physically the same, the attributes or the criteria must have changed if the intervention had any effect. These changes might have affected the incentives only lightly. Suppose for instance that insurance companies are now obliged to lay bare all exclusions of a contract, instead of, say, scattering them throughout the contract or having them footnoted in small print. If the customer discovers a condition that had escaped his notice, his incentives to accept the contract may change to any degree, and in particular so slightly that nudge 1 would hold unproblematically.

Having been satisfied with this property, we turn to the other, and now difficulties arise. If RECAP interventions are concerned with bounded rationality, this could only be with core biases unlike with the SMT case, but are they concerned with bounded rationality at all? On one reading of a RECAP instruction, companies will provide new information to the agents, so that nothing can be concluded on the latter's rationality or otherwise. On theoretical ground, it is impossible to exclude that they have applied some decent updating procedure, be it Bayes's rule or something else. The other reading, which is presumably Thaler and Sunstein's, is that companies will present the same information differently, that ideal consumers could make the translation by themselves, but that real consumers need assistance, because they fall prey to the biases just said. Which of the two readings is the more acceptable?

One argument for the informational reading is that RECAP will typically force companies to adduce extra information. For example, if there exist some legal means to cancel a tacitly renewable subscription beyond the relevant withdrawal period, one might expect RECAP to require the seller to explain them, and this will typically enlighten the ignorant consumer. However, this argument leads one to divide concrete RECAP arrangements into two groups, one satisfying the former reading, and the other the latter, depending on whether or not extra information is effectively added. The former group of arrangements is presumably the larger, but it is enough for Thaler and Sunstein's case if the latter group is non-empty, and this is also presumably the case.

An apparently stronger - and certainly more subtle - argument is that even if companies simply translate preexisting information in a more accessible way, this process requires technical knowledge that it would be unfair to regard as being required by rationality. Consider the apparently simple task of taking a loan from a bank and agreeing with it on an
amortization plan. In most countries, banks are legally required to apportion interest repayment and principal repayment in the amounts due by the customer, and they typically provide various amortization charts that differ in both this apportionment and the duration of the loan. This is an example of a RECAP translation. Now, there seems to be no irrationality involved in the fact that the borrower needs the bank's accountants to make the calculations. We agree that this is a slippery argument. What part of mathematics should be included in the mental equipment of a rational agent? Only simple arithmetic or also the use of the exponential function? Depending on whether one stops early or late along the slippery slope, one will prefer the first or the second reading of RECAP.

If we cannot reach any firm conclusion here, this in part because *framing* is the bias at work, and it is the most elusive of all. Here we meet the objection that framing has irregular effects and it may be difficult to base an intervention on its working. But there is an even more basic problem: even if RECAP involved bounded rationality, the latter would *be the object and not the means*; so RECAP would be, like withdrawal periods, only a case of nudge 3 without nudge 2. Suppose that the customers' idea of the loan was distorted by some framing effect and the banker now redescribes the loan conditions in such a way that this effect vanishes. It certainly helps the banker to be aware of the customers' typical misconceptions, but it is then knowledge of framing, not framing itself, which has an instrumental sense. Things would be different if the banker exploited a framing effect to make the new picture successful, but none of Thaler and Sunstein's examples suggests that RECAP should include manoeuvring in this way. We read it as being a straightforwardly corrective device.

As a fourth example, consider the across-the-board recommendation in *Nudge* to include a default option into the choice set. We must assume either that a preexisting element of the set is turned into a default, or that a new element is added with this property, and that the agent is well aware of the change. Otherwise, nudge 1 would be violated straightforwardly. Now, under these assumptions, the choice situation is altered to the following extent: *the agent has the choice between choosing and not choosing*, and if he makes the first (meta) choice, he will be faced with exactly the same (basic) choice set as before. Although the new

---

27 Normatively, they recommend it across the board for complex financial and they review the organ donation problem in this light; they could have also discussed class action, an area in which the matter of default options looms large. Class action systems are divided into those with opting out and those with opting in, and this distinction corresponds to two ways of setting the default option.
situation does not formally count as an increase of the initial (basic) set, the choice possibilities are enlarged, and we may conclude that the first condition of nudge 1 is met. The problem lies with the other condition that incentives should not change substantially. One objection is that the agent could take the default option to be recommended, and this could block him from choosing another option he approves of. Another, more troubling objection will be recognized once we have checked that nudge 2 holds.

Psychological biases play a definite role with default options. The intervention fights against the biases of availability, representativeness, and anchoring and adjusting. Gregariousness, which is a theoretically ambiguous phenomenon, may also play a role. But the sense in which the intervention puts bounded rationality in motion is different; what then matters is that it saves the agent the pain of deliberating about the choice objects. Our sketchy model does not take research costs explicitly into account, and thus does not pay justice to this important sense of bounded rationality. 28 But however construed, it should involve a change in incentives – not in the incentives to choose such and such option, which relate to the standard biases, but in the incentives to choose at all. For example, the more trustworthy the nudger is in the agent's eyes, the more inclined the latter will be to accept the default option blindly. Thus, the condition for nudge 1 is not met, and this is another case in which the first two senses fall apart. As a consolation, the nudge 3 property is satisfied. Default options are a nice example in which some biases (i.e., availability, etc) represent the obstacle to be removed and another fault (i.e., deliberative incompetence) serves as the instrument.

The table on the next page sums up the results of this section.

---

28 This is Simon's sense of bounded rationality rather than Kahneman and Tversky's. Nudge does not pay it full justice either.
### Cafeteria example

<table>
<thead>
<tr>
<th></th>
<th>Nudge 1</th>
<th>Nudge 2</th>
<th>Nudge 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Save More Tomorrow (SMT)</td>
<td>yes</td>
<td>dubious</td>
<td>dubious</td>
</tr>
<tr>
<td>- <em>ex ante:</em></td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>- <em>ex post:</em></td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Withdrawal periods</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Record, Evaluate, Compare Alternative Prices (RECAP)</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Default options</td>
<td>Choice set: yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>Incentives: no</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4. Back to the normative

The normative discussion initiated by Thaler and Sunstein is primarily geared at libertarian paternalism, with two overlapping lines of inquiry. The former questions Thaler and Sunstein's understanding of the two doctrines they claim to reconcile, i.e., liberalism and paternalism, as well as related conceptual issues in moral philosophy. The latter, which is more applied, addresses libertarian paternalism from the angle of nudging interventions. This line is actually closer to Thaler and Sunstein's approach to moral notions, which basically consists in studying examples. They argue that their privileged cases of nudges satisfy the requirements of both liberalism and paternalism, granting that, conversely, these examples help specify the ill-defined requirements. This is a reflective equilibrium methodology, and one may question the particular equilibrium they propose. The work of the last sections has some bearing on both lines of inquiry.
The first line has led to virtually unanimous rejection of libertarian paternalism. Thaler and Sunstein oppose the common view of paternalism by claiming that it does not necessarily entail coercion. What they keep from paternalism once coercion goes away is by and large the approval of interventions that aim at improving another party's welfare even when that party does not acquiesce in the change. By reflective equilibrium, the cafeteria example is supposed to confirm this and related intuitions of what paternalism consists of. A recurring objection in the literature that paternalism cannot be defined so widely and that coercion - although not an entirely clear idea - should somehow be included to make the relevant restriction.

Along the same abstract line of inquiry, another recurring objection is that Thaler and Sunstein conceive of liberalism also too broadly. Some criticisms amount to arguing that liberalism, in a genuine sense, requires a deeper notion of freedom than just freedom of choice, which is the only one considered by Thaler and Sunstein. Other criticisms amount to endorsing freedom of choice as a sufficient criterion of liberalism, while complaining that Thaler and Sunstein handle it inadequately. Both groups of criticisms feed on what may be called the anti-manipulation claim: liberalism permits interfering with the choices of another party only if that party has sufficient knowledge and understanding of the interference. Both groups of criticisms take for granted that Thaler and Sunstein fall prey to the anti-manipulation claim. Indeed, it seems to go without saying that bounded rationality can be exploited for a purpose only without the agents' full awareness of the intervention ("in the dark").

These are complex objections, and they may involve faulty intermediary steps even if they are basically sound (in particular, the assumption that interventions can exploit bounded rationality only "in the dark" is in fact dubious). The work of the previous sections provides a shortcut to the conclusion that Thaler and Sunstein have not reconciled liberalism and

29 "The second misconception is that paternalism always involves coercion" (2009, p. 11; already in 2003a and b). The "first misconception" they consider is that "it is possible to avoid influencing people's choices" (2009, p. 10).
30 See especially Hausman and Welch (2009).
31 See in particular Mitchell (2005) and Grüne-Yanoff (2012). The former critic reinforces the present liberal objection with a redistributive objection that we do not examine here (being intended for the less rational, nudging interventions would impose a cost on the more rational). This added objection connects with the idea of "asymmetric paternalism" in behavioural economics.
paternalism. However these doctrines are precisely defined, Thaler and Sunstein's case for their reconciliation depends on the claim that the nudge 1, 2 and 3 concepts are compatible, a factual claim that the last section has precisely rebutted. In the authors' view, nudge 1 carries the liberal component, nudge 3 the paternalistic component, and nudge 2 - which specifies the technology - the conciliatory component. Supposedly, by exploiting bounded rationality, an intervention can be so devised as to be liberal, in the nudge 1 sense, and paternalist, in the nudge 3 sense. It does not matter how Thaler and Sunstein use the moral terminology if it is shown that nudge 2 interventions are rarely, if ever, also nudge 1 and nudge 3. Admittedly, our shortcut is effective only as an ad hominem argument. It may be that liberalism and paternalism can be reconciled without referring to nudges. However, the critics have not given an absolute proof either that reconciliation was impossible, and for the main, their own arguments are also ad hominem.

The second line of inquiry is precisely concerned with how Thaler and Sunstein use the words, and whether they have gotten the right meanings by reflecting on the many examples they provide. Here again, the general opinion is negative. The more careful analysis of nudges in sections 2 and 3 unfortunately warrants this rebuttal. To consider only case, the examples of nudge 1 interventions do not enter as they should into a reflective equilibrium with the meaning of "respecting freedom of choice". Rather than suggesting a mild intervention, of the kind that would indeed preserve freedom of choice, they suggest a non-intervention, which would nullify the idea of acting while respecting it (and thus would in the end make liberalism an impotent doctrine). This is because, as we have seen, it is too stringent to impose the two conditions at the same time - not forbidding any option and not significantly changing the economic incentives. Section 2 has cast a doubt on the conjunctive requirement that section 3 confirms with the example of default options. Here, the choice set condition holds, but the incentive condition does not, and one would nonetheless think that the intervention preserves freedom of choice (arguably, it even reinforces it by allowing for the possibility of not choosing). To interpret "respecting freedom of choice" in terms of the choice set condition alone is not the solution, as it would allow for odd examples such as blackmailing or sexual harassment; it seems as if something should be added on incentives, though not in the way nudge 1 says it. ³² Thus, the examples fail their task of clarifying

³² The authors must have realized by themselves that the choice set condition was not sufficient. "Choices are not blocked or fenced off" was all they required in Sunstein and
"respecting freedom of choice", and the reflective equilibrium is not yet reached. We leave the intriguing semantic question open-ended, having only meant to reach this last claim.

It would be disappointing if the paper contributed only to simplifying and reinforcing the quasi-unanimous rebuttal of Thaler and Sunstein. Actually, we view their work as being of primarily technological, not ethical significance. Correspondingly, we view nudge 2 and 3 are more promising concepts than nudge 1, as they suggest new ways of intervention to public policy makers and those who advise them. At least when economic advising is concerned, public interventions are both justified and planned in terms of individual rationality schemes. These schemes may be rough or precise, but they have the overall effect of reducing the scope of interventions to those suboptimal situations, like market failures, in which irrationality arises only as the collective level. The simple idea behind nudge 3 are that rationality failures at the individual level also set a possible task for public interventions. They would affect individuals distributively, not collectively, as in education, which may offer an even better heuristics than what remains of Thaler and Sunstein's list of examples. The not simple idea underlying nudge 2 combined with nudge 3 is that one rationality failure can be used to counter the suboptimality effect of another, and again, some educative devices would illustrate that plainly. The new area of the public economics of bounded rationality should be explored only with an eye on the efficacy of interventions, and this may entail putting to their right place such teasing ideas as that of bounded rationality being turned from illness to remedy. The corrective purposes of nudge 3 can often be more naturally realized by enlightening the agents, or by providing them with self-control devices, which do not belong to the realm of bounded rationality. These purely individual explorations are anyhow just a first step, because individual rationality failures intermingle with collective rationality failures in the actual world, and public economics would be discharged of its obligations only if it could take both sets of failures into account.

---

Thaler (2003a, p. 1162), and the incentive condition appeared later in Nudge presumably because they were dissatisfied with this initial statement.
References


