Behavioural Public Policy (2018), 2: 1, 107–124
 doi:10.1017/bpp.2016.16

Rethinking nudge: not one but three concepts

PHILIPPE MONGIN*

CNRS and HEC Paris

MIKAËL COZIC

Université de Paris-Est-Créteil and Institut Universitaire de France

Abstract: 'Nudge' is a concept of policy intervention that originates in Thaler and Sunstein's (2008) popular eponymous book. Following their own hints, we distinguish three properties of nudge interventions: they redirect individual choices by only slightly altering choice conditions (here 'nudge 1'); they use rationality failures instrumentally (here 'nudge 2'); and they alleviate the unfavourable effects of these failures (here 'nudge 3'). We explore each property in semantic detail and show that no entailment relation holds between them. This calls into question the theoretical unity of nudge as intended by Thaler and Sunstein and most of their followers. We eventually recommend pursuing each property separately, both in policy research and at the foundational level. We particularly emphasise the need for reconsidering the respective roles of decision theory and behavioural economics to delineate nudge 2 correctly. The paper differs from most in the literature in focusing on the definitional rather than the normative problems of nudge.

Submitted 6 December 2016; accepted 14 December 2016

Introduction

Nudge is a concept of policy intervention that originates in Thaler and Sunstein's (2008) eponymous book and has disseminated from there to various areas of law, economics, philosophy and social theory.[1] Thaler and Sunstein (henceforth T&S) introduce it as a general category, allowing the intervening parties to be either public or private and to pursue any kind of

* Correspondence to: GREGHEC, CNRS and HEC Paris, 1 rue de la Libération, F-78350 Jouy-en-Josas, France. Email: mongin@greg-hec.com

1 *Nudge* (Thaler & Sunstein, 2008) is intended for a wide audience and should be supplemented by more academic work published both earlier and later by the authors, whether jointly or separately. However, because it is the core source, we will primarily refer to the book, here cited in the paperback, slightly expanded edition (*Nudge*, Thaler & Sunstein, 2009).

interest. However, they specifically emphasise, and argue for, nudges by public authorities that mean to increase the welfare of the population – in short, *welfare-promoting nudges* – and the lively discussions prompted by their work are mostly concerned with this class of interventions. Today's policy research has witnessed a collective effort to substitute traditional welfare economics with a more appropriate theoretical basis for public policies, and nudge is a major inspiration for this endeavour. However, even when it is so restricted, nudge can be understood in more than one sense, and this paper is about this semantic diversity. We hope to contribute to the current trend in policy research by analysing the concept more fully than is usually done.

According to one sense, a nudge is a policy intervention for redirecting an agent's choices by very slightly altering their choice conditions so that the interference is kept to a minimum. Having this sense in mind, T&S contrast welfare-promoting nudges with traditional public policies, which typically rely on bans, commands or heavy manipulations of choice incentives. In another sense, nudge is a policy intervention that reaches its objective by taking advantage of the rationality failures of the choosers. T&S strongly emphasise cognitive and practical limitations to decision making, as evidenced by today's behavioural economics, and their book is also a survey of these findings. Since traditional public policies often, if not always, instrumentally rely on rational responses from individuals, a new contrast emerges between these policies and welfare-promoting nudges; arguably, the latter have a more satisfactory empirical basis than the former. However, T&S do not simply view individual rationality failures as a new technique of intervention. They also believe that these failures diminish welfare and should be removed, or at least alleviated, by suitable interventions, which they also describe as being nudges; this establishes still another sense of their concept.

In sum, nudge can mean: (1) an intervention that interferes with the choice conditions minimally; (2) an intervention that uses rationality failures instrumentally; and (3) a welfare-promoting intervention that tries to reduce the negative effects of rationality failures. We will unimaginatively say 'nudge 1', 'nudge 2' and 'nudge 3', respectively, for these properties. This paper will elaborate on these three definitions and, once they are more precisely fixed, investigate their logical relations. An earlier version complemented this semantic analysis with a critical discussion of T&S's concrete examples of nudges. We have dispensed here with this step not only for the sake of brevity, but also because we believe that the semantic analysis has independent interest. The original idea of nudge is intrinsically equivocal, as the three meanings above testify, and no doubt in part for this reason, it has spread out anarchically in behavioural economics and policy research. Writers in these fields often claim the label for a variety of proposed interventions without

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

explaining what this means for them; others justify its wide-ranging application by entirely idiosyncratic definitions. Not surprisingly, the lists of nudges that circulate contain clearly irrelevant items. The success of this promising idea coincides with a semantic laxness that it is perhaps time to restrain.

Besides innumerable applications, the nudge idea has given rise to a reflective literature in social theory and philosophy, and a word is needed to locate our paper in this corpus. Most contributors here are also, and in fact primarily, concerned with *libertarian paternalism*, a social ethics conception that T&S defend at the same time as welfare-promoting nudges. In the two authors' work, this conception antedates the terminology and, to some extent, the very idea of nudge.[2] Their longstanding interest is to reconcile libertarianism (in the sense of respecting the individual's freedom of choice) and paternalism (in the sense of giving priority to welfare improvement over the individual's spontaneous will). A major argument in *Nudge* is that welfare-promoting nudges implement libertarian paternalism, and thus provide evidence that this is a feasible doctrine despite its inherent tension. Unlike most of the reflective literature, our paper investigates neither this alleged connection, nor libertarian paternalism in any other way; it entirely shifts the focus away from normative to semantic analysis. To our knowledge, only Bovens (2009) and Hansen (2016) investigate nudge distinctly from libertarian paternalism, and only the latter compares possible definitions as we do here. This is not to suggest that joint discussions of nudge and libertarian paternalism are unhelpful for our purposes, and the comparison section will comment on some semantic ideas taken from these normative discussions.

The starting point in *Nudge*

We extract two main suggestions from *Nudge*.

"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 Econs" (2009, p. 8).

These two sentences enunciate meanings for nudge 1 and 2, respectively. We currently ignore the fact that T&S refer to a "factor" or an "aspect" of the

2 The word 'nudge' does not appear in T&S's work before the book. However, several of the interventions proposed there belonged to their earlier papers (Sunstein & Thaler, 2003a and 2003b) and were critically discussed at the time (see Mitchell, 2005).

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

choice conditions rather than an intervention directly; this is a secondary equivocation to be discussed later. The pair of sentences can be understood differently depending on how one reads "our definition" in the second. This expression may either refer to what comes next in the second sentence or to what was previously said in the first. Depending on this grammatical choice, either nudge 1 or nudge 2 is the real definition, with the other property being only derivative. Whichever solution one chooses, the two properties need to be logically related, but we will discuss this problem only after fixing more precise meanings for each separately.

Nudge 1 and its two conditions

Regarding nudge 1, there are two conditions: (i) *not forbidding any option* and (ii) *not significantly changing the economic incentives*.[3] Condition (i) is best construed in terms of *physically defined options*. Not only does this interpretation give more bite to (i) than if one takes the options to be subjectively perceived, but it also helps delineate (ii) by contrast; that is, (ii) will concern the manipulation of non-physical – notably financial – aspects of the options. We take the "not forbidding" requirement at face value, thus permitting the set to remain the same or to be enlarged, and excluding all reductions, unlike Hansen (2016, p. 167), who argues that some are compatible with nudge. The possibility of a larger set plays a critical role below.

The analysis of condition (ii) is more troublesome. On the one hand, it seems natural to call upon standard economics to decide what "changing the economic incentives" means, and this suggests interpreting this in terms of changes in financial constraints or more primitive changes in either prices or personal incomes. The interventions usually considered in welfare economics change prices through indirect taxes or subsidies and personal incomes through direct taxes or subsidies. An advantage of this very standard interpretation is that it turns (ii) into an objective condition. On the other hand, it is so restrictive as to make (ii) virtually powerless. To borrow a counterexample from Hansen and Jespersen (2013) and Hansen (2016), an electroshock therapy would satisfy the incentive condition for nudge. Hence, the temptation of restating (ii) without the "economic" adjective. T&S themselves warrant this move in a footnote they append to the first sentence: "Some of our nudges do, in a sense, impose cognitive (rather than material) costs, and in that sense alter

3 The authors have initially contented themselves with the first condition. "Choices are not blocked or fenced off" was all they required in Sunstein and Thaler (2003a, p. 1162). Only in *Nudge* and later work do they envisage the second condition.

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

incentives. *Nudges count as such … only if any costs are low*" (2009, p. 6, emphasis in original).[4] However, to extend (ii) in this way would defeat the purpose of treating nudge 1 objectively. Cognitive costs are not directly observable apart from in some experimental contexts, and to assess them indirectly by changes in observable behaviour poses an obvious risk of circularity. Once these two sources of evidence are excluded, there only remain introspection reports and verbal testimonies, which are not entirely reliable sources to use.

We suggest an intermediary solution, which keeps the "economic" adjective, but construes it less narrowly than was first proposed. By drawing on basic decision theory rather than basic economics, we will consider not only (1) a *physical set of options* and (2) *financial constraints*, but also (3) *beliefs* and (4) *preferences*. By definition, a change in economic incentives will consist of a change in one or more of these four determinants of decisions, granting that changes in the option set must be increasing so as to preserve compatibility with (i). Thus, we introduce two more channels for an intervention besides standard economic constraints; one may induce new choices also by acting on preferences and beliefs. These two channels can be explored empirically either by the specialised techniques of decision theory or, more naively, by the means of folk psychology. Beliefs and preferences are the target attitudes of mind reading, in the folk-psychological sense, and they can be contrasted with cognitive costs, which mind reading is not concerned with. This brief sketch is enough to suggest that it is easier to glean information on beliefs and preferences than on cognitive costs.

The introduction of beliefs and preferences is compatible both with a classical rationality modelling, as in expected utility theory, and non-standard modelling in which no probability appears, preferences may not be ordered and adding options may violate revealed preference conditions. Thus, we do not sever possible connections with nudge 2. Some would recommend *rank-dependent utility theory* (RDU) as an intermediary ground, but there are other theoretical possibilities, and we do not need to make a choice here.[5]

Nudge 2 and the problem of identifying rationality failures

In the sentence above, T&S express nudge 2 by contrasting "Humans" with "Econs," a pleasant but obscure allegory. Fortunately, they soon translate the allegory into drearier, but more usable language. They do not mean

4 The issue of cognitive versus material costs surfaces again in Sunstein (2014, p. 38).

5 RDU generalises expected utility by allowing distortions of cumulative probabilities and, in some versions, accommodates the endowment effect discussed below; see Wakker's (2010) up-to-date presentation. Oliver (2013) includes RDU in the corpus of behavioural economics.

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

different populations; rather, they view each individual as hosting both a Human and an Econ.[6] They offer two renderings for this internal division. The first – see T&S (2009, p. 19) and, more emphatically, Sunstein (2014, Ch. 1) – exploits a famous construction of recent psychology, which separates two modes of cognitive functioning, also described as 'systems'. System 1 is supposedly quick, semi-automatic, not fully conscious and likely to make mistakes; System 2 is supposedly slow, reflective and deliberative, fully conscious and less susceptible to mistakes. By a major tenet of this theory, each system has a consistent mode of operation and works essentially independently of the other. Each system is better adapted than the other to specific tasks, but they can nonetheless compete or cooperate on the same task. The second theoretical rendering – see T&S (2009, Ch. 1–3) – exploits a list of rationality failures borrowed from current behavioural economics.

Thus, there are two possible meanings for nudge 2, either as an intervention that brings about a response from System 1 and none from System 2 or as an intervention that instrumentally relies on rationality failures taken from the list. Two-system theory fits well with the picture of a Human and an Econ coexisting within the same individual, but it has the two drawbacks of being contentious and not easily applicable in practice. As it turns out, the list is T&S's effective tool of analysis, and as in the bulk of the literature, we will restrict attention to it.[7] There are three broad groups in it: 'biases and blunders', 'temptation' and 'following the herd'. We briefly consider them in turn, having two guiding questions in mind: is each item really a rationality failure? And does behavioural economics really handle each item better than decision theory can do? Notice the two questions, although closely related, are distinct. In the same spirit, Hausman and Welch ask: "Why shouldn't these factors be regarded as interferences with rational choice rather than as rational determinants of choice?" (2010, p. 126). We unfold Hausman and Welch's question by distinguishing what pertains to rationality as broadly thought of and what pertains to the disciplinary comparison of behavioural economics with decision theory.

The third group – 'following the herd' – hardly passes the test of our two questions. On the one hand, not all gregarious behaviour involves a rationality failure. For example, bank runs and financial crashes can be rationalised as

6 By contrast, *asymmetric paternalism* considers interpersonal differences in responding to interventions (Camerer *et al.*, 2003). More on this disanalogy in Mitchell (2005).

7 Heilmann (2014) and Selinger and Whyte (2011) use two-system theory, but these are fairly uncommon examples. This theory raises intense controversies in current psychology. Compare the sharp objections in Kruglanski and Gigerenzer (2011) and Gigerenzer (2010) with the reply in Evans and Stanovitch (2013).

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

second-order reactions to the other participants' initiatives. No criterion exists to screen off these rationalisable cases from those in which no individual interest can justify the gregarious behaviour. *Nudge* states alleged examples of this behaviour and treats them as if they automatically escaped rationality, but these are dubious classifications. Some experiments show that reluctant tax payers are more sensitive to messages like 'Most people in this area pay their taxes' than to messages like 'By paying your taxes, you fund the costs of public utilities'. It is unclear whether this finding uncovers a genuine rationality failure (the former message suggests a warning that the latter does not) and even whether it involves gregariousness in the first place (this should not be confused with a sense of reciprocity and cooperation).[8] On the other hand, the disciplinary comparison is not what it should be. The behavioural economics of gregariousness is marred by an identification problem, which the tax example illustrates, whereas economics can provide a formal definition and handle some cases, in particular by game theoretic tools. These converging arguments suggest excluding the third group altogether – this is not a big blow to *Nudge*, because this group is under-represented compared with the other two.

The second group – 'temptation' – passes the double test much better, but cannot be endorsed unreservedly. Giving in to temptation is often, though not always, coincidental with being *time inconsistent*. That is, agents acting against their own will are often, though not always, also reneging on an earlier commitment not to act this way. Now, time inconsistency is taken to be a rationality failure across the board, and the disciplinary comparison is this time what it should be. Standard decision theory has little to say on time inconsistency that is not simply a normative indictment, and behavioural economics has developed an insightful model – *hyperbolic discounting* – to account for it descriptively (e.g. Laibson, 1997). However, time inconsistency is difficult to ascertain empirically because of a possible confounding with *a change in information*. Agents who seem to renege on previous commitments may in fact react to what they have recently learned, and they may even rationally do so. Here, decision theory takes revenge because it can offer well-structured analyses in terms of Bayesian revision or even more general models, whereas behavioural economics has little to contribute beyond the finding that real agents often violate Bayesianism.

We will also express reservations on the first group – 'biases and blunders' – with its classic list inherited from Kahneman, Slovic and Tversky (1989): anchoring and adjustment, availability, representativeness, overconfidence,

8 T&S (2009, p. 67) cite an early Minnesota study on tax compliance. More telling evidence can be found in Hallsworth *et al.* (2014).

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

loss aversion, status quo and framing. One may wonder why the first three items constitute rationality failures at all. Anchoring and adjustment can be a normatively commendable procedure for selecting an option, as some formal algorithms illustrate, and availability and representativeness are so loosely defined that it is equally easy to include them in, or exclude them from, individual rationality.[9] Concerning the disciplinary comparison, it is true that these items were first identified by Tversky and Kahneman (1974) in connection with empirical difficulties of Bayesian decision theory. This is not to say, however, that decision theory per se cannot handle them, and even less that behavioural economics handles them better.

The next three items raise different issues. Presumably, the argument for treating overconfidence as a rationality failure is that it induces the individual to neglect available evidence or make faulty reasoning. But viewed in a different light, overconfidence is a *character trait*, and as such neither rational nor irrational. Decision theory can then try to absorb it into its utility apparatus, as it has done with risk attitudes. We see no argument for claiming that loss aversion and status quo involve rationality failures.[10] The concept of rationality does not preclude that the agents' ends should be stated in terms of differences from some reference level, as status quo implies. Nor does it exclude the notion that the agents' evaluation may be influenced by what they already possess, as loss aversion implies. Whether behavioural economics handles these two items better than decision theory is a different matter. The former has brought out telling evidence that agents often compare satisfactions in terms of differences and by paying attention to endowments, and the latter has been sluggish in taking these messages on board. To some extent, the disciplinary comparison depends on how one locates those decision theories, like RDU, which empirically supersede expected utility theory.

Framing is a strange outlier in the list. By definition, a framing effect occurs when equivalent descriptions of choice conditions induce different choices from the agents. Framing is an *effect*, which is compatible with a diversity of outcomes, rather than a *bia*s, which normally produces outcomes in a fixed direction. Also, Tversky and Kahneman's well-known suggestion that biases result from misapplying *heuristics* cannot plausibly concern framing. Heuristics are particular cases of rules, and unless other biases are also present, one cannot see what rule framed agents would be following. The

9 For Gigerenzer (2010), availability and representativeness illustrate the cheap use of labels instead of models, which he calls "one-word explanations." The same label often permits explaining one phenomenon and its contrary.

10 Notice that status quo and loss aversion are sometimes identified and sometimes kept distinct, a further cause of embarrassment. In a rare effort, Brenner *et al.* (2007) try to disentangle them.

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

famous epidemics examples in Tversky and Kahneman (1981) are unrepresentative since they involve the status quo or loss aversion biases (e.g. see Frisch, 1993). Besides these important differences, and perhaps in connection with them, framing is a better candidate than any other bias to the status of a rationality failure, and it is also most recalcitrant to decision theoretic treatments – at least none is in view thus far. At long last, the two questions seem to elicit clear, positive answers. However, as with time inconsistency, there is a problem of empirical identification. The subjects of a framing experiment may not agree with the equivalence postulated by the experimenter, and they can moreover have serious reasons for this disagreement. Kahneman belatedly came to recognise this major difficulty (in the introduction to Kahneman and Tversky, 2000).

Taking stock of the analysis, we have found nudge 2 more difficult to define than nudge 1. The semantic work is entangled here with an unsettled debate as to what counts as a rationality failure and with a complex diagnosis of how behavioural economics compares with decision theory when the latter is interpreted with some subtlety and not polemically reduced to constrained optimisation and expected utility theory. Only time inconsistency and framing provide an unquestionable basis for nudge 2; all other cases are open to interpretations.

Logical relations between nudge 1 and nudge 2

We see only one way of logically deriving nudge 2 from nudge 1. Suppose an intervention leaves an agent's option set and economic incentives *exactly* unchanged – a supposition permitted by the definition of nudge 1. Now suppose the intervention nonetheless succeeds in altering the agent's choice. The conclusion seems inescapable that the decision process lacked rationality: either the agent had sufficiently good reasons for the initial choice and thus was wrong to revise it, since nothing changed in the choice conditions; or the agent did not have sufficiently good reasons for the initial choice, which points to a different lack of rationality. When rationality is understood in the standard economics sense of constrained optimisation, the argument is even more briefly put: optimisers stay where they are when constraints do not change.

However, we cannot yet conclude that the intervention is nudge 2. The argument shows that a rationality failure is *involved*, not that it is *used instrumentally*, as this property requires; for example, an accidental framing effect might have taken place. There is an even more sweeping objection to make. When the option set increases or the economic incentives undergo minor changes – two cases that are permitted by (i) and (ii) – the argument does not apply anymore. However slight the changes, it may be rational to revise the initial choice. One may try to extend the range of the argument by claiming that slight changes in the choice conditions normally deliver slight changes in the

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

choices, so that a large change in the latter would signal a lack of rationality. But this is an implausible claim; rational choice is compatible with strong discontinuities, as a lexicographic ordering of options illustrates.

Now consider the opposite logical direction; in other words, from nudge 2 to nudge 1. There is one clear derivation from *framing*, since this effect does not require choice conditions to change, and nudge 1 permits this complete stability. But framing is just one example of nudge 2, and the others do not support the derivation. Thus, an intervention based on loss aversion needs to change the endowment, hence the option set, and this change may not be increasing, as against (i). An intervention based on overconfidence will change the preferences or beliefs relative to some options, hence the economic incentives in our sense, and this change may not be light, as against (ii). The more biases one takes to be relevant to nudge 2, the more difficult it is to derive nudge 1. In sum, no entailment holds between the two concepts at the desired level of generality.

Nudge 3 and its relation to nudge 2

We may now clear up a secondary problem in T&S's definition of nudge. Explicitly, they define it as an "aspect" or a "factor" of the choice conditions (or "choice architecture," as they like saying), not as a kind of policy intervention. To connect the two senses, one will have to say that a nudge *intervention* consists in using a nudge *factor* to exert an influence on the choices. This seems to be too roundabout, given that policy interventions are the focus of attention and, like most readers, and indeed Sunstein (2014, p. 17; 2015, p. 417) himself in recent work, we prefer defining nudges directly as being interventions and ignoring the notion of nudge as a causal factor. There is another reason for this simplification. T&S repeatedly argue for welfare-promoting nudges by saying that the individuals are already influenced by other nudges in their ordinary choices. This *inevitability argument* has attracted much criticism in the normative discussion. As the objection goes, it is not the same for an individual to be influenced by a factor in the choice conditions when someone uses this factor for a purpose and when the factor is just there without anyone in particular being in control; see, for example, Hausman and Welch (2010) and Grüne-Yanoff (2012). Without delving further into this important debate, we may point out that using the same word for the intervention and its underlying mechanism can only foster confusion between the two cases, and this is a good reason for avoiding this language.[11]

11 Sunstein's current position may sound paradoxical because he emphatically reiterates the inevitability argument while redefining nudges as being interventions. However, the argument can

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

More importantly, when T&S deal with nudge interventions, they allow the intervening party to pursue any objectives of its own, thus taking the benevolent objective of welfare promotion to be a mere particular case. Marketing research has applied the nudge concept to some of the branding, packaging and advertising policies by which the consumer product industry hopes to push up its sales. The properties of nudge 1 and 2 are applicable to these interventions, but this is not the case with nudge 3, which captures a subclass of welfare-promoting interventions.

We define nudge 3 interventions as those that *counteract the rationality failures affecting individuals' decision processes*. T&S do not introduce this property as explicitly as the first two, but it clearly underlies most of their concrete recommendations. We sketchily review these recommendations, mentioning for each what rationality failures it is meant to counteract. (1) The introduction of self-commitment devices opposes the tendency to time-inconsistent reversals of preferences and procrastination. (2) The introduction of withdrawal ("cooling off") periods opposes thoughtless choice and biases such as overconfidence. (3) The imposition of disclosure practices on businesses and administrations opposes thoughtless choice and the framing effect. (4) The introduction of default options or forced choices, depending on the area, opposes choice overloading, the tendency to procrastination and biases related to inertia such as loss aversion. One may worry that T&S tend to enlarge their initial list of rationality failures when they approach concrete examples, but we do not push this point here.

Rather, we capitalise on these examples to explore the conceptual dimensions of nudge 3. We distinguish two of them, one regarding *the possible ways an intervention counteracts rationality failures* and the other regarding *the nature of the improvement brought about by the intervention*. To spell out the first dimension, an intervention can be preventative or only mitigating; in other words, it can preclude rationality failures from occurring in the decision process or permit their occurrence while limiting their impact on this process. Thus, withdrawal periods and forced choices are preventative, whereas default options are only mitigating. A third possibility that is *not* illustrated here, but enters the definition of nudge 3, is to let the failures occur and act on the choices themselves; for example, by penalising or compensating transfers. This *ex post* intervention, which is in the spirit of traditional welfare economics, involves neither prevention nor mitigation.

still be defended on the view that innumerable previous interventions have shaped today's choice conditions.

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

To spell out the second dimension, an intervention can be only remedial or only corrective or both at a time. By a remedial intervention, we mean one that counteracts the rationality failure in the given choice situation, and by a corrective intervention, we mean one that teaches the agent something generally usable about rationality failures. Thus, the imposition of withdrawal periods is merely remedial and to penalise poor decisions *ex post* would be merely corrective. The imposition of disclosure practices can have both properties. RECAP requires credit card companies to draw a clear separation between interest rates and fees in their customers' statements. This should not only influence the way customers use their credit cards with their company, but also teach them a diffidence rule they could remember in other dealings, such as when they take a mortgage loan from a bank.

The involvement of nudge 3 with rationality failures makes it superficially close to nudge 2, and some common formulations – like 'nudges draw on the findings of behavioural economics' or 'nudges trade on bounded rationality' – erase the distinction. This lowers theoretical standards dramatically, since it is so much easier to reconcile nudge 1 indiscriminately with nudges 2 or 3 than with just one – and *a fortiori* the two – of them. There is a clear difference between the two properties: a nudge 2 intervention uses the failures instrumentally for *whatever objective the intervention pursues* and a nudge 3 intervention counteracts them by *whatever means are deemed relevant*. Not only are these concepts distinct, but they bear no entailment relations. To check that in one logical direction, think of information and persuasion efforts, which satisfy nudge 3, but may not involve any instrumental use of rationality failures, hence may not satisfy nudge 2. The reverse entailment is also blocked, even if one only considers welfare-promoting nudge 2 interventions. Think of a public health (e.g. vaccination) campaign based on a framing effect. From Tversky and Kahneman (1981) and later confirmatory evidence, the campaign should be more successful with a 'positive' frame (emphasising the success rate) than with a 'negative' frame (emphasising the complementary failure rate). This intervention employs a bias for a welfare-related purpose without trying to counteract it.

The previous example is somewhat extreme because to counteract framing, in whatever sense we take for this, can only nullify an intervention based on framing. This is not necessarily the case with other rationality failures. For example, an intervention can counteract one effect of hyperbolic discounting while strategically using another of its effects. Thaler and Benartzi (2004) have argued for a new pension saving scheme precisely in this way. Their recommended pension scheme, Save More Tomorrow (SMT), makes increases in future savings coincidental with future increases in incomes. Arguably, this feature exploits the psychology of hyperbolic discounters. These agents believe they will save appropriate amounts in the future, but do not do so when the

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

time comes, and they are the more easily deceived since the actual saving decision is to be made in a more remote future. By starting with relatively low amounts and suitably delaying the increases, one takes advantage of their tendency for self-deception to counteract their threatening time inconsistency.

Also, one rationality failure can serve to fight another. Here are possible examples: employ overconfidence against the endowment effect and the tendency to inertia, availability against overconfidence and framing against any other bias. Some of these examples are sketched in *Nudge* and explored more thoroughly elsewhere. They capture what seems to be a major heuristic for developing the idea of nudge (i.e. the combination of nudges 2 and 3 to devise innovative welfare-promoting interventions). In principle, nudge 1 has no role to play in this heuristic. Plainly, it neither entails nor is entailed by nudge 3, and it can only be an interesting coincidence if the three properties are met on the same intervention.

Comparisons

Our semantics of nudge can briefly be compared with others in the literature. Bovens (2009) defines a nudge as "a manipulation of people's choices via the choice architecture, i.e., the way in which the choices are presented to them." In our terms, this amounts to restricting nudge to be nudge 2 and, in effect, to the employment of framing. This restriction fosters the claim that nudges are *manipulative*, a claim that loses plausibility when other rationality failures are taken into account. For instance, it is hard to see why the offer of self-committing devices should always be manipulative; think of a self-committing device being entirely rigid and offered just one period before the future decision, so that the time lag effect involved in SMT is not present. The following definitions capture richer concepts of nudge.

Here is Hausman and Welch's (2010, p. 126): "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."[12] In our terms, this takes nudge interventions to satisfy the conjunction of nudges 1, 2 and 3 together, a maximally demanding conception of nudge.

Oliver's (2013, p. 4–5) definition also compounds properties: "For an intervention to be classified as a nudge it needs to be liberty preserving, rely on the

12 We read a similar definition in Selinger and Whyte (2011, p. 926), though these writers generally emphasise the more limited conjunction of nudges 1 and 2.

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

automatic, reflexive responses of those targeted and not involve overly overt methods of persuasion, not significantly change economic incentives and has to redesign the choice context according to the findings of behavioural economics." This roughly equates nudge with the conjunction of nudge 1 (interpreting "liberty preserving" in the sense of (i)), nudge 2 (using a System 1 criterion of "reflexive responses") and a non-overtness property we do not include in nudge 2 (again, because this property primarily concerns framing). Nudge 3 is not mentioned, but this remains a demanding conception.[13]

Hansen's (2016, pp. 158 and 170) definition is difficult to quote without reviewing its complex supportive argument. Roughly speaking, it minimally consists of nudge 2 compounded with nudge 3 and, in an expanded variant, with a further property that substitutes for nudge 1. This only requires that the interventions "work independently of" decreasing the option set or significantly changing the incentives (here widely defined). Accordingly, Hansen permits interventions that instrumentally depend on both rationality failures and nudge 1-violating changes. The unity of nudge becomes less problematic and its scope correspondingly enlarged.[14]

Conclusions

Starting from T&S's theoretical hints for nudges 1 and 2 and their concrete examples for nudge 3, we made these properties more precise and explored their logical relations. Hopefully, this will have brought some order to the promising but still confused ideas of the book. Our framework has also facilitated comparisons within the secondary literature, an exercise that should be extended beyond the sample presented here. One of our conclusions is that no logical relation holds between the properties at the proper level of generality. There is no doubt that T&S meant nudge to be a unitary concept, and the definitions we quoted generally aim at capturing this intention. To reconcile it with our negative logical finding, two distinctive moves suggest themselves. One consists in *redefining one or more of the three properties*. This is well illustrated by Hansen's replacement of nudge 1 by a less demanding variant. Alternatively, one could restrict nudge 2 to framing, so as to make nudge 1

13 Oliver (2015) usefully emphasises a property of welfare-promoting nudges we have abstracted from here; i.e. that it does not aim at remedying negative compositional effects or negative externalities. This is now well recognised by Sunstein (2014, p. 28), who contrasts "internalities," the sole object of nudge, with collective rationality failures.

14 McQuillin and Sugden (2012, p. 560) and Grüne-Yanoff (2012, p. 639) are not so explicit about what they mean by nudge, but we can liken their concepts to nudge 2 conjoined with nudge 3. Like Qizilbash (2012) and still others, they are largely concerned with an issue not addressed here – what kind of preferences the intervening party should attribute to the party intervened on.

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

derivable from nudge 2, or limit nudge 1 to exactly unchanged choice conditions, so as to make nudge 2 derivable from nudge 1. The fact that redefinition strategies tamper with T&S's *wording* is inessential, since they primarily aim at endowing nudge with more theoretical substance, and it is enough for this project to keep in touch with T&S's *heuristics*. We have followed a similar interpretative line here. However, these strategies call for more specific objections. Hansen's property makes nudge compatible with reductions of the choice set and significant changes in incentives, and this could make nudge difficult to separate from traditional interventions. And the above restriction strategies involve too much of a loss of content to be pursued.

Another response is to keep the definitions and make unitary sense of nudge *in a non-logical, (i.e. factual) way*. If this line is taken, there must be a sufficiently large set of examples satisfying the three properties, or at least two of them, for nudge to remain a unitary concept, and this could be decided only by reviewing concrete examples.[15] Here we can at least make a methodological claim in advance. There is of course nothing wrong in defining a concept by a collection of properties that are only factually related. Scientific fields like biology and astronomy often proceed in this way; think of the standard species concepts in the former field or – a lesser known example – of the planet concept in the latter. However, the role of concepts so defined appears to be limited to facilitating description and classification. There are no deep theoretical stakes involved in the definition of a fish versus a mammal, or a planet versus a dwarf planet. These contrasts vary through time with the discovery of new objects that are difficult to classify, and when this happens, the theoretical structure of the field remains untouched. As this quick comparison suggests, one should not expect too much from a nudge concept that would exhibit some factual overlap between nudges 1, 2 and 3 and no more semantic unity than that. It would certainly help classify existing interventions and help discover interventions not yet thought of, but it could hardly be the basis for a theoretical revolution in policy research.

There is still another line that consists in *giving up the unitary perspective and pursuing each property in isolation*. A refined nudge 1 concept can help delineate what a light intervention is; there is nothing in traditional welfare economics that exactly captures this idea. A refined nudge 2 concept would open up new avenues for both policy research and more foundational work. Against the received idea that behavioural economics is unorthodoxy, we argued that decision theory (rather than economics) could appropriate alleged biases by treating them as psychological determinants of rational choice. More generally,

15 For a preliminary sketch, see Mongin and Cozic (2014).

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

the boundaries of the two fields in policy research must be considered anew, and this task will succeed more easily if one pursues it having only nudge 2 in mind. Last but not least, a refined nudge 3 concept can decisively extend the realm of public policies; policy research has too long ignored the fact that rationality failures called for remedies and corrections. Here, again, there is an advantage in developing the concept in isolation. Actually, Jolls and Sunstein's (2006) "debiasing through law" programme was like a separate implementation of nudge 3 before nudge officially entered the stage, for better or worse. They proposed that the law should respond to rationality failures not only by adjusting legal norms accordingly (e.g. by raising safety requirements to take overconfidence into account), but also by "operating directly on the boundedly rational behavior and attempting to help people either to reduce or to eliminate it" (2006, p. 201) (e.g. by requiring firms to provide vivid evidence of difficult cases and clear warnings that their products may be dangerous). In our terms, this amounts to exploring nudge 3 without entangling it with either nudge 1 or even nudge 2.[16]

Acknowledgements

The authors have benefitted from useful comments obtained during seminars or conferences in Trento, Paris, Nashville, Lyon, Saint-Denis, Rotterdam and Genève. Many thanks to Adam Oliver for checking the present manuscript.

Disclaimer

This is a shorter and much revised version of Mongin and Cozic (2014).

References

Bovens, L. (2009), 'The Ethics of Nudge', in T. Grüne-Yanoff and S. O. Hansson (eds), *Preferences Change: Approaches from Philosophy, Economics and Psychology*, Berlin: Springer, 207–219.

Brenner, L., Y. Rottenstreich, S. Sood and B. Bilgin (2007), 'On the Psychology of Loss Aversion: Possession, Valence, and Reversals of the Endowment Effect', *Journal of Consumer Research*, **34**: 369–376.

16 Sunstein's recent work is in line with the unitary heuristics of *Nudge*, but continues to emphasise nudge 3 or even broader ideas, possibly at the expense of other properties of nudge. When he writes: "Nudges can counteract biases (such as unrealistic optimism) without exploiting anything" (2014, p. 38), he focuses on nudge 3 interventions that are not also nudge 2.

https://doi.org/10.1017/bpp.2016.16 Published online by Cambridge University Press

Camerer, C., S. Issacharoff, G. Loewenstein, T. O'Donoghue and M. Rabin (2003), 'Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism', *University of Pennsylvania Law Review*, **151**: 1211–1254.
Evans, J. S. and K. E. Stanovitch (2013), 'Dual-Process Theories of Higher Cognition: Advancing the Debate', *Perspectives on Psychological Science*, **8**: 223–241.
Frisch, D. (1993), 'Reasons for Framing Effects', *Organizational Behavior and Human Decision Processes*, **54**: 399–429.
Gigerenzer, G. (2010), 'Personal Reflections on Theory and Psychology', *Theory and Psychology*, **20**: 733–743.
Grüne-Yanoff, T. (2012), 'Old Wine in New Casks: Libertarian Paternalism Still Violates Liberal Principles', *Social Choice and Welfare*, **38**: 635–645.
Hallsworth, M., J. A. List, R. D. Metcalfe and I. Vlaev (2014), 'The Behavioralist As Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance', NBER WP No. 20007.
Hansen, P. G. (2016), 'The Definition of Nudge and Libertarian Paternalism: Does the Hand Fit the Glove?', *European Journal of Risk Regulation*, **7**: 155–174.
Hansen, P. G. and A. M. Jespersen (2013), 'Nudge and the Manipulation of Choice. A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy', *European Journal of Risk Regulation*, **4**: 3–28.
Hausman, D. and B. Welch (2010), 'To Nudge or Not to Nudge', *Journal of Political Philosophy*, **18**: 123–136.
Heilmann, C. (2014), 'Success Conditions for Nudges: A Methodological Critique of Libertarian Paternalism', *European Journal for Philosophy of Science*, **4**: 75–94.
Jolls, C. and C. R. Sunstein (2006), 'Debiasing through Law,' *Journal of Legal Studies*, **35**: 199–242.
Kahneman, D., P. Slovic and A. Tversky (eds) (1989), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.
Kahneman, D. and A. Tversky (eds) (2000), *Choice, Values and Frames*, New York: Cambridge University Press.
Kruglanski, A. W. and G. Gigerenzer (2011), 'Intuitive and Deliberate Judgments Are Based on Common Principles', *Psychological Review*, **118**: 97–109.
Laibson, D. (1997), 'Golden Eggs and Hyperbolic Discounting', *Quarterly Journal of Economics*, **112**: 443–477.
McQuillin, B. and R. Sugden (2012), 'Reconciling Normative and Behavioural Economics: The Problems to Be Solved', *Social Choice and Welfare*, **38**: 553–567.
Mitchell, G. (2005), 'Libertarian Paternalism is an Oxymoron', *Northwestern University Law Review*, **99**: 1245–1277.
Mongin, P. and M. Cozic (2014), 'Rethinking Nudge', *HEC Paris Research Paper No. ECO/SCD-2014–1067*.
Oliver, A. (2013), 'From Nudging to Budging: Using Behavioural Economics to Inform Public Sector Policy', *Journal of Social Policy*, **42**: 685–700.
Oliver, A. (2015), 'Budging, Shoving, and Budging: Behavioural Economic-Informed Policy', *Public Administration*, **93**: 700–714.
Qizilbash, M. (2012), 'Informed Desire and the Ambitions of Libertarian Paternalism', *Social Choice and Welfare*, **38**: 647–658.
Selinger, E. and K. P. Whyte (2011), 'Is There a Right Way to Nudge? The Practice and Ethics of Choice Architecture', *Sociology Compass*, **5**: 923–935.
Sunstein, C. (2014), *Why Nudge? The Politics of Libertarian Paternalism*, New Haven: Yale University Press.
Sunstein, C. (2015), 'The Ethics of Nudging', *Yale Journal on Regulation*, **32**: 413–450.
Sunstein, C. and R. H. Thaler (2003a), 'Libertarian Paternalism Is Not an Oxymoron', *The University of Chicago Law Review*, **70**: 1159–1202.

Sunstein, C. and R. H. Thaler (2003b), 'Libertarian Paternalism', *American Economic Review*, **93**: Papers and Proceedings, 175–179.
Thaler, R. H. and S. Benartzi (2004), 'Save More Tomorrow: Using Behavioral Economics to Increase Employee Savings.' *Journal of Political Economy*, **112**: 164–87.
Thaler, R. H. and C. Sunstein (2008), *Nudge*, Yale: Yale University Press.
Thaler, R. H. and C. Sunstein (2009), *Nudge*, New York: Penguin Books.
Tversky, A. and D. Kahneman (1974), 'Judgment Under Uncertainty: Heuristics and Biases', *Science*, New Series, **185**(4157): 1124–1131.
Tversky, A. and D. Kahneman (1981), 'The Framing of Decisions and the Psychology of Choice', New Series, *Science*, **211**(4481): 453–458.
Wakker, P. (2010), *Prospect Theory for Risk and Ambiguity*, Cambridge: Cambridge University Press.