Economics and Philosophy, **34** (2018) 243–266 © Cambridge University Press
doi:10.1017/S0266267118000032 First published online 26 April 2018
cambridge.org/eap

TOWARD A FRAMEWORK FOR SELECTING BEHAVIOURAL POLICIES: HOW TO CHOOSE BETWEEN BOOSTS AND NUDGES

TILL GRÜNE-YANOFF,[*] CATERINA MARCHIONNI[†], MARKUS A. FEUFEL[‡]

Abstract: In this paper, we analyse the difference between two types of behavioural policies – nudges and boosts. We distinguish them on the basis of the mechanisms through which they are expected to operate and identify the contextual conditions that are necessary for each policy to be successful. Our framework helps judging which type of policy is more likely to bring about the intended behavioural outcome in a given situation.

Keywords: behavioural policy, nudge, boost, heuristic, decision-making
JEL codes: B40, D04, D60

1. INTRODUCTION

Behavioural policy – the application of insights from behavioural research to public policy – has found widespread acceptance in recent years (e.g. Oliver 2013; Shafir 2013; Bhargava and Loewenstein 2015; Chetty 2015). Many governmental and non-governmental organizations currently advise or decree the use of behavioural insights when designing

[*] Royal Institute of Technology (KTH), Division of Philosophy, Brinellvägen 32, 10044 Stockholm, Sweden. Email: gryne@kth.se

[†] Centre for Philosophy of Social Science, Department of Political and Economic Studies, University of Helsinki, Unioninkatu 40A, 00014 Helsinki, Finland. Email: caterina.marchionni@helsinki.fi.

[‡] Technische Universität Berlin, Department of Psychology and Ergonomics, Division of Ergonomics, Marchstr. 23 (MAR 3-2), 10587 Berlin, Germany. Email: markus.feufel@tu-berlin.de. URL: http://www.awb.tu-berlin.de/.

interventions to achieve specific policy goals.[1] Proponents of behavioural policies often adopt a pragmatic perspective that stresses the dependence of both the success and the ethical acceptability of different behavioural interventions on the policy context (e.g. Chetty 2015; Soll *et al*. 2015; Sunstein 2016). In other words, motivated by concerns about extrapolation and external validity, they claim that the choice of a policy tool should ultimately depend on considerations specific to the context of intended application (for earlier philosophical arguments of context-dependence see Steel 2008; Cartwright and Hardie 2012). This recognition has so far not led to an approach that systematically includes context-dependence in planning, testing and implementing policies, however. Given that testing a behavioural policy, for example in a field experiment, often already amounts to its partial implementation, it is recommendable to make optimal use of both theoretical and empirical resources *before* deciding how to intervene in a specific context.

Within the domain of non-incentivizing, non-coercive policies, in this paper we focus on two types of behavioural interventions: *nudges* and *boosts* (Grüne-Yanoff and Hertwig 2016; Hertwig and Grüne-Yanoff 2017). As a first approximation, nudges steer people towards a particular behaviour by creating environmental conditions that trigger a given heuristic strategy. In contrast, boosts change behaviour by fostering people's decision-making competences in a given environment. Our analysis has two goals. First, it aims to clarify a central distinction between nudges and boosts by differentiating each policy type on the basis of the *mechanism* through which each is supposed to affect behaviour. Mechanistic knowledge is often useful for extrapolating evidence of effectiveness, that is, for making reliable inferences about whether the behavioural intervention will be effective in a different context or in the long run (e.g. Grüne-Yanoff 2016). Second, based on these mechanistic differences, it aims to develop a framework that can inform the choice between these two types of behavioural policy in the run-up to empirical test and implementation.

We will proceed as follows: Section 2 reviews existing definitions of nudges and boosts. Section 3 proposes two mechanistic models that refine existing distinctions and illustrates each model with two sets of cases. Section 4 analyses the two mechanistic models to derive the contextual

[1] Including the *Behavioural Insights Team* in the UK (ps://www.gov.uk/government/organisations/behavioural-insights-team), and the Executive Order *Using Behavioral Science Insights to Better Serve the American People*, signed by president Obama on 15 September 2015 (https://www.whitehouse.gov/the-press-office/2015/09/15/executive-order-using-behavioral-science-insights-better-serve-american). More generally, the recent OECD report surveys 159 uses of such policies from 23 member states (OECD 2017).

conditions that each type of policy has to satisfy in order to be effective. Section 5 concludes.

2. DEFINITIONS

Grüne-Yanoff and Hertwig (2016) have recently proposed a distinction between two types of behavioural policies, which they associate with two theoretical positions in the behavioural sciences – *nudges* with the *Heuristics and Biases* program (Tversky and Kahneman 1986; Kahneman and Tversky 1996) and *boosts* with the *Fast & Frugal Heuristics* program (Gigerenzer *et al*. 1999).

Nudges and boosts are similar in that they seek to change behaviour without substantially changing material incentives and without coercing the agent through legal prohibitions or mandates. Furthermore, they both assume that people use a limited set of heuristics to make decisions and that the success or failure of these heuristics depends on properties of the decision environment (Thaler and Sunstein 2008; Grüne-Yanoff and Hertwig 2016). By contrast, nudges and boosts differ both in their ethical implications and in how they are hypothesized to affect behaviour. In this paper, we bracket issues pertaining to the ethical assessment of the policies. Instead, we focus on mechanistic differences between the two policy types and derive the conditions under which each is effective in relation to a specified policy goal such as increasing gym attendance or saving rates. We also abstract from considerations of the political economy of behavioural public policy by assuming that the policymaker is a 'benevolent planner' (e.g. Schubert 2017). This is not because we think ethical or political considerations are unimportant or uninteresting. Bracketing them out, however, is instrumental to our purpose of distinguishing the two types of policy solely on the basis of the mechanisms through which they are supposed to operate.

In their *Nudge* book, Thaler and Sunstein (2008: 6) define nudges as:

> any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid.

Within the domain of non-incentivizing and non-coercive behavioural policies, this definition is very inclusive. The only substantial constraint it imposes is that nudges affect behaviour by intervening on the *choice architecture* – that is, the properties of the environment in which the choice is made. Thaler and Sunstein specify the above definition further by arguing that nudges operate by *instrumentalizing* the effect of a biasing heuristic. Status quo bias is an example of such a biasing heuristic for them: it leads to systematically predictable inferior decisions. Thus, status quo bias has a powerful influence on people's behaviour, and 'that power can be harnessed' (Thaler and Sunstein 2008: 8). Starting from the assumption that people use a limited set of heuristics to make decisions (an assumption that it shares with the boost approach), the nudge approach instrumentalizes these cognitive limitations to influence behaviour. This harnessing strategy is one of the core ideas of the nudge approach.

By contrast, boosts have been characterized by their 'goal of expanding (boosting) the decision maker's set of competences and thus helping them to reach their objectives' (Grüne-Yanoff and Hertwig 2016: 156). According to this definition, boosts are distinct from nudges in at least two ways. First, boosts seek to realize this goal by expanding agents' set of competences; they aim to overcome human cognitive limitations rather than instrumentalizing or harnessing them. Second, boost interventions do not exclusively target the choice environment, but often target the agents' heuristic repertoire directly.

Therefore, the contrast between nudges and boosts that is relevant to the goal of evaluating their effectiveness is as follows: a nudge intervenes on people's choice environment and harnesses a certain heuristic to bring about a specific behavioural change; a boost intervenes on people's heuristic repertoires and expands that repertoire to bring about a specific behavioural change.

We acknowledge this way of drawing the distinction does not fit all available definitions. For example, some authors think that all behavioural policies are nudges (Sunstein 2016). The situation is further complicated by a lack of a widely accepted definition of nudges: different authors have proposed varying characterizations (e.g. Bovens 2009; Hausman and Welch 2010; Rebonato 2012; Heilmann 2014; Hansen 2016; Mongin and Cozic 2017). Given that boosts have been defined in contrast to nudges, these ambiguities might also affect the boost concept.

Our paper improves the understanding of the distinction between these two types of behavioural policies.[2] In particular, it expands on and further clarifies the conceptual distinction proposed in Grüne-Yanoff and Hertwig (2016) and Hertwig and Grüne-Yanoff (2017) by explicitly modelling the underlying mechanisms of the respective policy types. In fact, although Grüne-Yanoff and Hertwig (2016: 163) acknowledge the importance of mechanisms, they do not perform a mechanistic analysis of nudges and boosts and largely focus on the coherence between policy approaches and the theories of bounded rationality that motivate them. Hertwig and Grüne-Yanoff (2017: 17–19) discuss the distinct 'causal

[2] The purpose of our analysis is to capture the core distinction between these policy types. Therefore, our definitions do not aim to replicate the preferred list of policy interventions of any specific author.

pathways' through which these policies affect behaviour, but draw on the psychological notion of a 'cognitive architecture' to explain this difference. Furthermore, neither of the above papers analyses the implications that a mechanism-based distinction has for ex-ante judgements of the applicability and effectiveness of the two types of policy.

3. DISTINGUISHING BOOSTS AND NUDGES BY MECHANISMS

The effectiveness of a policy can be assessed empirically, for example by performing randomized controlled experiments. The celebrated *Behavioural Insight Team* has adopted this approach for assessing the effectiveness of behavioural interventions. It has become increasingly clear, however, that knowledge of the mechanism whereby a policy operates greatly helps in planning and designing the right empirical trials, in correctly interpreting their results, and most crucially, in exporting a policy to other populations and settings (Cartwright 2010; Ludwig *et al*. 2011; Sampson *et al*. 2013; Clarke *et al*. 2014; Grüne-Yanoff 2016). Two different policies aiming to achieve the same goal might exhibit similar *effect sizes* in trials, but might operate through different mechanisms that require different contextual conditions for their successful operation. Disregarding this mechanistic information might lead to the wrong conclusion that both policies are equally applicable in different contexts.

Current philosophy of science characterizes *mechanisms* broadly as systems of causally interacting parts and processes, which – under certain conditions – predictably produce one or more effects (e.g. Craver and Tabery 2016; Glennan 2016; Marchionni 2017). In the case at hand, the relevant mechanisms link policy interventions to agents' behaviour. The link between these components is mediated by the agents' decisions and the environment in which these decisions are taken. In particular, the interventions with which we are concerned here are mediated by decisions that are assumed to result from the use of a limited set of *heuristics*. The behavioural effect of these heuristics in turn depends on properties of the decision environment. Therefore, the relevant components of the mechanisms are (i) the *heuristic repertoire R*, which refers to the set of heuristics individuals have at their disposal, (ii) the *environment E* and (iii) the resulting *behaviour B*. Clearly these components are but a narrow selection of the *context of intended application* of a behavioural intervention, which also includes political, institutional and other sociocultural features. These simple representations nevertheless allow us (i) to draw a clear-cut distinction between boosts and nudges and (ii) to provide a systematic analysis of policy effectiveness across contexts, that is, particular instantiations of the environment, heuristic repertoire and behaviour. If at a later point, further details are considered relevant for the analysis, these simple representations can be easily expanded.

The main distinction we posit here relates to the way in which a choice environment *E* is connected to the application of a particular heuristic in repertoire *R*. According to the Heuristics & Biases literature, which guides our analysis of the nudge mechanism, heuristics tend to be stable: type of environment *E* always triggers the same heuristics and hence leads to predictable behaviour. The older Heuristics & Biases literature often considers heuristics to be cognitive illusions and sees them as analogous to visual illusions (Tversky and Kahneman 1986; Kahneman and Tversky 1996), which cannot be adjusted (e.g. in the Müller–Lyer illusion, we 'see' that one line is longer than another, even if we know that this impression is false). Similarly, the more recent literature often considers the heuristics responsible for behavioural deficits to be located in System 1, which 'operates automatically and quickly, with little or no effort, and no sense of voluntary control' (Kahneman 2011: 20). In particular, this 'System 1' is believed to be 'not readily educable' (Kahneman 2011: 417), which implies that it is hard to change with interventions. Irrespective of what the underlying conceptual models are, most Heuristics & Biases proponents seem convinced that 'our ability to de-bias people is quite limited' (Thaler cited in Bond 2009). Based on the view that heuristics are stable, changing the environment *E* appears as a promising avenue toward predictable behavioural change. Consequently, in the nudge mechanism, the intervention aims at influencing behaviour *B* by modifying the agents' choice environment *E*.

By contrast, according to the Fast & Frugal Heuristics literature, which guides our analysis of the boost mechanism, the choice of heuristics from repertoire *R* is to some extent controlled by the agent. This requires the agent to be equipped with various competences, including those needed to identify the problem at hand and the goals pursued, as well as to select the heuristic from the repertoire that (is believed to) further the goal. Hence, training the agent in new decision tools can be used to change her behaviour, even when the environment *E* does not change. Although the adaptive use of one's heuristic repertoire might not be universal, the Fast & Frugal Heuristics program contends that:

> An automatic rule is adapted to our past environment without a present evaluation as to whether it is appropriate. It is simply triggered when the stimulus is present . . . . The flexible rules, in contrast, involve a quick evaluation of which one to use . . . . Gut-feelings may appear simplistic, but their underlying intelligence lies in selecting the right rule of thumb for the right situation. (Gigerenzer 2007: 49)

The literature contains ample examples of behavioural domains in which agents predictably switched from *automatic* to *flexible* decision rules (Pachur and Hertwig 2006; Volz *et al*. 2006; Mega *et al*. 2015). Take, for example, the *recognition heuristic*, which suggests choosing the option one recognizes over less familiar or unknown options to identify the largest among a number of cities. Clearly, such a heuristic is successful only in environments in which recognition differentiates between alternatives and is associated with the relevant selection criterion. Researchers found that agents are sensitive to this association when applying the heuristic:

> people appear to frequently overrule recognition information in an environment in which there is little to no relationship between recognition and the criterion. Indeed, we found that in such an environment, the use of the recognition heuristic was restrained. (Pachur and Hertwig 2006: 993)

In the boost mechanism, the assumed relation between environment E and heuristic repertoire R is thus different from the nudge mechanism. Instead of triggering a particular heuristic, the environment provides informational cues, which inform the agent's selection of a heuristic. For instance, if cues in the environment do not match the recognition heuristic (e.g. there is no relationship between recognition and the criterion), then another heuristic is chosen. Because here heuristics are selected by the agents rather than triggered, it is possible to bring about changes in behaviour B by way of training agents on how and when to use novel heuristics. This is why we characterize boosts as intervening on the agents' heuristic repertoire R rather than on the environment.

On the basis of these considerations, we derive two stylized mechanistic models of nudges and boosts (Figure 1).[3] In both models, behaviour B is caused by the application of heuristic h from repertoire R in environment E.

Let the heuristic function h represents the causal influence of R and E on B and the subscripts $_n$ and $_b$ represent the boost and the nudge model, respectively. In nudges, heuristics are assumed to be *stable* so that changes in environment E lead to predictable changes in behaviour B:

$$B_n = h_n(E)$$

To capitalize on this relationship, nudge interventions target and modify the environment, as indicated by the red arrow on the left of Figure 1.

By contrast, in the boost mechanism, the chooser selects heuristics adaptively depending on their fit with a given environment, so that

$$B_b = h_B(R, E)$$

[3] We use the term *model* to highlight the fact that they are simplified representations that leave out many features of what a realistic description of the mechanisms behind heuristic decision-making would presumably entail. Many of these features however are likely to be common to both the nudge and the boost mechanism. Our purpose here is to focus attention on their differences.

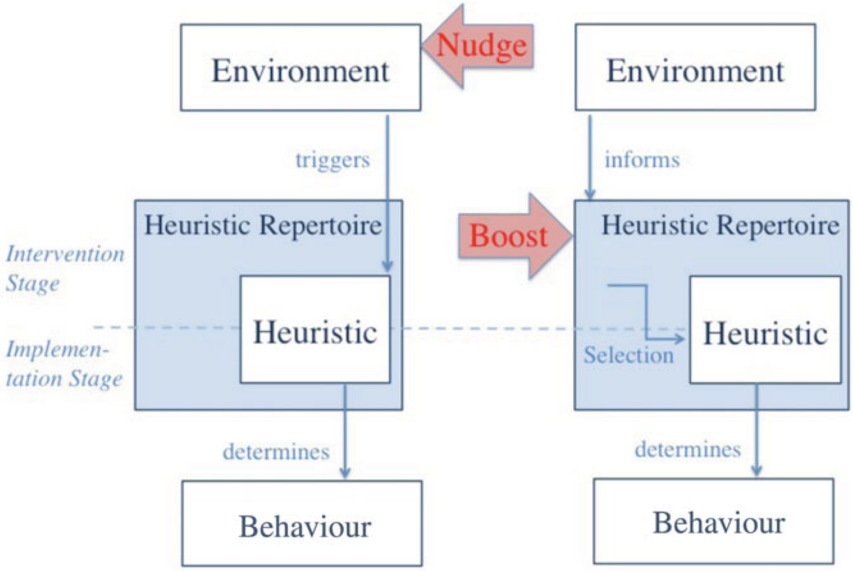

FIGURE 1. Mechanisms of nudge and boost interventions (Colour online)

Given that the behavioural outcome *B* is mediated by the agent's selection of heuristics from *R*, boost interventions target the heuristic repertoire *R* as indicated by the red arrow on the right of Figure 1.

Consequently, the models distinguish nudges and boosts in two ways: by the mechanisms they assume to be at work to produce behaviour (i.e. heuristics as stable or flexible factors) and, more importantly for our purposes, by the component of the mechanism the intervention targets (i.e. environment *E* or heuristic repertoire *R*).

Note that here we do not ask: 'Which model is generally true?' but rather: 'Which intervention is effective in which context?' (see also Chetty 2015). In practice, this means keeping open the possibility that the two models are not competing representations of heuristic-based decision-making, but apply to different kinds of heuristics (Chow 2015; Polonioli 2016). Finding out whether a heuristic-produced behaviour can be more effectively intervened upon through a nudge or a boost is ultimately an empirical question – but asking and evaluating this question requires an understanding of the conceptual distinction and of its underlying mechanisms.

To illustrate the difference between the nudge and the boost mechanisms in more concrete terms, we now consider examples of behavioural interventions that aim at producing the same outcome but do so in different ways.

3.1. Financial behaviour

Policymakers and experts in the financial domain have been concerned that many people lack sufficient financial literacy to successfully plan lifetime consumption or run a small business. Different behavioural interventions have been proposed with the goal of improving financial behaviour.

Save More Tomorrow (SMT) is a behavioural intervention designed to increase people's retirement savings. Its design is based on viewing decisions about retirement savings as trade-offs between current and later (retirement-age) consumption and aims to influence these decisions by changing one of the options in the trade-off (Thaler and Benartzi 2004; Benartzi and Thaler 2013). That is, instead of asking people to choose a trade-off between consumption now or later, the policymakers ask people to choose a trade-off between consumption in the near future (say, a year from now) and later. The rationale behind this intervention is that people tend to systematically overvalue the immediate present over the future, thus reducing willingness to save for retirement now. The theory of *hyperbolic discounting* (Loewenstein and Prelec 1992) captures this finding but also suggests that modifying the timing of options may be used to reduce this effect. The SMT program also takes into account people's inertia (increases in savings are automatic) to ease them into self-control restrictions (by projecting them into the future). Various observational studies show that this intervention is effective. Among people who said that they could not afford a cut in pay now, 78% joined the programme (see Thaler and Sunstein 2008).

An alternative policy targeting financial behaviour trains people in the use of *simple rules of thumb* (SRT) for making financial decisions (Drexler *et al*. 2014). For example, people were trained to keep their money in two separate drawers (or purses) for their business and household, respectively, and to transfer money from one drawer to the other only when accompanied by an explicit 'I Owe You' note. This simple rule of thumb for structuring revenues had significant and substantial effects on business practices, reporting errors and revenues, in particular in low-skilled individuals (Drexler *et al*. 2014). Notably, training a control group in standard accounting practices did not have comparable effects.

Both SMT and SRT have been shown to be effective interventions to facilitate far-sighted financial behaviour. And although they pursue closely related goals, they do so through different pathways. SMT intervenes on a feature of the environment – the temporal location of the saving choice. By shifting it into the future, SMT seeks to harness the relatively low discounting rates between two events that are located at a sufficient distance in the future, as implied by hyperbolic discounting. SMT expects that the change, under the hyperbolically discounting heuristic, will cause an increase in saving choice. It thus uses knowledge about a feature of people's heuristic repertoire to change the choice environment such that people are likely to adjust their behaviour accordingly. According to our distinction, SMT is a nudge.

SRT, in contrast, is a boost because it intervenes in people's heuristic repertoire by teaching them a new rule of thumb. The rule helps choosers to interpret the environment in new ways (according to the household/business account distinction), and will only be applied if matching information cues (i.e. separate sources of income and expenditure that match this distinction) can be found in the environment. To the extent that the training is successful, it expands the agents' heuristic repertoire so that they exhibit better business discipline and are more likely to succeed in their businesses.

3.2. Gym attendance

Policymakers and experts in the health domain have been concerned that many people do too little exercise and therefore are more likely to suffer from cardiovascular and other diseases. Again, different behavioural interventions have been proposed with the goal of increasing gym attendance.

Providing gym membership by default is such a policy. Typically, gym memberships are 'opt-in', since an employee might get a subsidy for membership but must choose to become a member to use this subsidy. The *default-setting* (DS) policy changes this to an 'opt-out': you are automatically enrolled, but if you explicitly choose not to be a member, you might use the subsidy for other services. The effectiveness of such policies has been shown in domains like saving for retirement (Beshears *et al.* 2009) but it has also been applied with the goal of increasing the exercise regime. For example, as part of a broad regime of health-improving measures, the Cleveland Clinic has implemented such an opt-out system and reports substantial weight loss and reduction in blood pressure amongst its employees (Klein 2011).

Alternatively, *Temptation Bundling* (TB) teaches people a strategy to overcome self-control problems by coupling 'instantly gratifying "want" activities (e.g., watching the next episode of a television show, checking Facebook, eating an indulgent meal) with engagement in a "should" behaviour that provides long-term benefits but requires the exertion of willpower (e.g., working out, completing a paper review)' (Milkman *et al.* 2013: 1–2). Specifically, the experimenters recommended that the subjects allow themselves to enjoy a number of desirable audio novels *only while exercising*, thus making their gym attendance more tempting. In comparison to a control group that was just given a bookstore voucher of the same value as the audiobooks (0.94 gym visits/week), the subjects using TB showed a higher weekly gym attendance rate (1.42 visits/week).

Both TB and DS aim at increasing gym attendance rates. As in the previous case, the two interventions do so through distinct pathways. TB aims to enrich the agents' heuristic repertoire by training them in the use of a new heuristic. This heuristic helps them interpret their environment in a novel way ('Am I facing a self-control problem with respect to the gym or not?'). If the chooser detects a positive cue in this dimension, she reacts by applying the bundling rule. This in turn affects her behaviour, making her visit the gym more often. Thus, according to our model, TB is a boost.

DS in contrast intervenes in the environment by changing the default setting of a choice situation. Many experiments have shown that people tend to choose the default option more often, be this for reasons of inertia, loss aversion or recommendation effects (Thaler and Sunstein 2008; Beshears *et al*. 2009). DS uses knowledge about these tendencies to change the environment in such a way that people are likely to choose the desired option. According to our model, DS is a nudge.

4. MODEL ANALYSIS: CONTEXTUAL CONDITIONS FOR THE EFFECTIVENESS OF NUDGES AND BOOSTS

So far, we have described how nudges and boosts affect the three components of our model: environment E, heuristic repertoire R and behaviour B in a given context. For each component, we now derive the conditions that must be satisfied for nudges and boosts to change behaviour via the posited mechanisms.

For the analysis, we think of boost and nudge interventions as two-stage processes. In Figure 1, these stages are separated by the dashed blue line. In the *intervention stage*, people's judgement and decision-making is affected by the policy. In the *implementation stage*, the judgement and decision is implemented into actual behaviour.

Concerning the conditions that must be satisfied, let us briefly consider our examples from Sections 3.1 and 3.2. At the intervention stage, the Save More Tomorrow (SMT) and Default Setting (DS) nudges intervene in the choice environment by altering the trade-off between future prospects and by resetting the default, respectively. The assumption is that, despite these changes, the intervention will trigger the same heuristic – hyperbolic discounting and the tendency to accept the default

– as before. If the assumption of stable heuristics were not met, the effectiveness of these nudge interventions would be jeopardized. Thus, *trigger stability* is a necessary condition for nudges.

At the intervention stage, the Simple Rules of Thumb (SRT) and Temptation Bundling (TB) boosts intervene in the agent's cognitive setup by training her to use new rules for structuring and processing information in the environment. Although SRT or TB need not necessarily involve a reflective judgement of what the rational decision is – for TB it is enough to train the agent to structure the environment in a way that is conducive to a better use of her temptations and impulses – the success of a boost crucially depends on successfully teaching/learning the new heuristics. Thus, failures of boost interventions include a lack of motivation or skill on the part of the agent or pedagogical failures for difficult-to-teach heuristics. Thus, the agents' *motivation* and the *teachability* of heuristics are necessary conditions for effective boosts at the intervention stage.

At the implementation stage, agents in a population may differ with respect to their level of competence and/or the heuristics they have at their disposal. This poses a problem for nudges, because nudges aim to change behaviour by harnessing people's heuristics, and such a harnessing must focus on the dominant heuristic in the population. A wide variety of employed heuristics in the population diminishes the effectiveness of nudges' one-size-fits-all solution and therefore requires a high degree of *homogeneity in people's heuristic repertoires* to yield reliable results. Boosts, in contrast, aim to enrich people's set of heuristics, and therefore remain untroubled by a high degree of heterogeneity in people's heuristic repertoires. A similar problem for nudges results from adaptations to an intervention, that is, when agents compensate for its behavioural effects or get used to it when it is applied repeatedly over time. Such *temporal effects* may interfere with the intended behavioural change.

For boosts, knowledge of *selection criteria* – when and when not to apply a heuristic – is a prerequisite at the implementation stage. Furthermore, stable selection patterns are threatened in situations where people do not have adequate *resources* to apply a taught heuristic, or where the heuristic contradicts prevailing values. We summarize these conditions in Table 1 (where a ■ indicates that a specific condition is required and a – that it is not) and then proceed to elaborate on each condition in turn.

4.1. Trigger stability

The Heuristics & Biases programme has documented many persistent cognitive biases, mainly through laboratory experiments (Tversky and

Conditions	Nudges	Boosts
Intervention Stage		
Trigger stability	■	–
Agent motivation	–	■
Teachability of heuristics	–	■
Implementation Stage		
Homogeneity of heuristic repertoire	■	–
Temporal effects	■	–
Selection criteria	–	■
Individual/Environmental resources	–	■

TABLE 1. Summary of conditions and impact on the effectiveness of nudges and boosts.

Kahneman 1974; Kahneman *et al*. 1982; Haselton *et al*. 2005). However, even if the existence of a bias is established, it is an additional question whether it will be stable when a nudge is implemented.

Let's consider the Save More Tomorrow (SMT) intervention discussed in Section 3.1. This nudge works if people show stable preferences related to hyperbolic discounting, that is, overvaluation of benefits obtained in the present over similar or even higher benefits available in the future. However, if the agent were aware that what matters for the decision is her view in a year's time she might adjust her evaluation and apply a different heuristic. In such cases, changing the timing of the options will not result in the predicted behavioural reaction. Similarly, if the agent interpreted the intervention as an attempt to manipulate her, she might seek to avoid the manipulation by changing her reactions (for evidence of these kinds of *reactance*, cf. Brehm and Brehm 1981; Campbell 2007). Similar complications might arise with respect to the example of Default Setting (DS). For example, 'in contexts that involve less effort, and clearer antecedent preferences, default rules are less likely to stick, and disclosures might make them less sticky still' (Sunstein 2016: 118). Also, if existing defaults are greater attractors than other options, because they are more salient, carry a signal better, or because they are less offensive, reactance to changes in default might result (Grüne-Yanoff 2016). In other words, when switching defaults as a policy intervention, one is not guaranteed that the same number of people will stick with the new default as they did with the old.

The goal of a nudge is to change the environment in such a way that the application of a given heuristic, which is assumed to be stable across environments, leads to a defined behavioural outcome. Hence, what needs to be established is that the triggering relationship between the environment and a heuristic is stable. If the relationship is unstable, policymakers cannot hope to reliably change behaviour by altering the environment. *In other words, nudge interventions require trigger stability between planned changes in the environment and known heuristics*. By contrast, since in our model of boosts heuristics are selected by the agent rather than triggered, the stability between the environmental trigger and the behaviour is not required for a boost to work. As we will see, boosts require a different sort of stability, namely sufficient (environmental) resources so that the agent is able to select the right heuristic from the heuristic repertoire.

4.2. Agent motivation

By definition, boosts aim at changing the agent's heuristic repertoire and therefore require the agent's active participation at the intervention stage. If the agent is either not able or not motivated to participate, it is unlikely (though not impossible) that a boost changes behaviour (see also Grüne-Yanoff and Hertwig 2016). Because ability is something to which a boost can in principle be calibrated but motivation is not, we examine each condition separately. We focus on *motivation* in this subsection and turn to *teachability* in the next.

Even the most inquisitive student will be bored or distracted by disorganized teaching. Thus, *motivation* clearly has a component that can be addressed by good pedagogical design. Motivation to learn and practice is, of course, a large topic that cannot be comprehensively discussed here. However, we indicate some conditions that might affect an agent's motivation. First, an agent's motivation depends on having a goal that she wants to achieve such as better health, financial stability or improved self-control. If she does not conceive of the intervention as a relevant means for her goal, a boost is less likely to be effective because she will see no reason to learn or undergo training in the first place. Second, even if an agent sees an intervention as a means to achieve her goals, she might not see the need to learn new heuristics because she believes in her current strategy and does not see the added benefit of an alternative one. Third, even if people are motivated to adjust their current strategy, they may not be motivated to learn unless they are convinced that the boost provides them with effective means to achieve their goals (e.g. thousands of purported 'dietary rules' are peddled – often by agents with less than sincere motives – and the success of most of these rules is very uncertain at best). In other words, if the level of expected effort appears higher than the expected benefit from learning, people will not be motivated to learn.

Thus, in case agents do not see the added benefit of learning new strategies to achieve their goals, boosts tend to be less effective than nudges. Of course, motivation to learn may be increased by appropriate information about the pros, cons and effectiveness of the boost before training. Sometimes, lack of motivation may even be addressed with specific boosts (or nudges). Nevertheless, even the best training interventions and preparatory information might fail to motivate people, for example, because they suffer from inertia or keep procrastinating. *In sum, in contexts where motivation is lacking, nudges tend to be more effective than boosts because a change in the environment does not require the agents' active participation*.

4.3. Teachability of heuristics

Motivated agents might fail to learn a heuristic because of pedagogical failures. Such failures might lie either in the design of the heuristic itself or in the way it is taught. For example, most people will be challenged to learn how to extract power roots without paper and pencil. Tools that require more memory and processing capacity than a person can muster are examples of such design failures. Less extreme examples include the difficulty of teaching low-skilled individuals standard accounting practices. As Drexler *et al*. (2014) showed, training small-business owners in standard accounting practices had no effect in improving reporting practices and business success, while simple rules of thumb training had a significant effect.

The Fast & Frugal Heuristics programme has provided evidence that simple, easy-to-grasp rules can be used to make sufficiently accurate judgements compared to more complicated procedures (Gigerenzer *et al*. 1999). The benefits of such decision-making strategies have been observed in a variety of settings, ranging from food consumption, to the stock market to online dating (van der Linden 2011). Thus, boosts are most effective in domains where simple and effective rules of thumbs can be identified or simple ways exist to teach a more complicated strategy. *Hence, when teachability is low, nudges tend to be the better choice*.

The second issue concerns good pedagogy. Here again, the relevant literature is extensive. We want to mention only one contextual factor that strikes us as important. As we have defined them, boosts aim to teach the right application of simple heuristics, which must be trained in repeated applications in order to be learned. An important feature of training success is sufficient and prompt feedback. In most training situations (such as schools, universities or PhD supervision), feedback is provided by the teacher. Since boosts are intended for large populations, not for individual training sessions, the feedback should be provided automatically. In fact, many successful boosts include such feedbacks. For example, people trying out Temptation Bundling (TD) might see their exercise time go up, or at least find it easier to regularly visit the gym. Similarly, people who have taken a first lesson in simple rules of thumb for running small businesses might see that applying the rules positively affects their business behaviour.

We do not believe that feedback is a necessary condition for the effectiveness of boosts in the strict sense, however. In the short run, Simple Rules of Thumb (SRT) may be effective even without the user being aware of its function. But without positive feedback, people have no reason to continue using a heuristic in the longer run. For example, it is plausible to assume that people would eventually discontinue using separate drawers for family and business accounts, unless they came to understand that this helps them achieve better financial discipline and ultimately better business outcomes. Also, feedback is not a sufficient condition for learning. For example, outcome feedback does not make people's tendency to underestimate rare events go away (Hertwig *et al.* 2004). However, there is evidence that feedback helps people learn to use their heuristics adaptively (Rieskamp and Otto 2006). *In general, feedback will help to create confidence in the effectiveness of a boost, which in turn will increase the likelihood that people persist and maintain motivation to learn and apply it*.

The previous three conditions concern requirements for effective boosts and nudges at the intervention stage. When the corresponding conditions are satisfied, nudges will trigger heuristics systematically and reliably and boosts will successfully enrich people's heuristic repertoire. However, this does not ensure that either intervention will change behaviour in the intended direction. For this to happen the heuristic triggered or the newly learned heuristic must reliably affect the behaviour of at least the majority of the targeted population in a particular direction. That is, there should be some positive net effect: for example, that more people show the behaviour than before, or – when there is a quantifiable behavioural outcome, like a retirement contribution – that the average outcome has changed in the desired direction. This leads to the discussion of the implementation stage of Figure 1, where heuristics or behavioural tendencies produce actual behaviour.

4.4. Homogeneity of heuristic repertoire across individuals

We start by analysing the conditions for the effectiveness of nudges at the implementation stage. Nudges typically affect the choice environment for a whole population based on the so-called *equal incompetence assumption*, which 'treats all ... actors in all situations as if they were equally predisposed to commit errors of judgment and choice' (Mitchell 2002: 2).

Thus, the behavioural triggers that nudges exploit do not, and typically cannot, discriminate between individuals who may have different heuristics at their disposal or simply differ in their levels of competency. Let us illustrate the problem of heterogeneity with an example from the Default Setting (DS) intervention. Beshears *et al.* (2009) describe the effects of resetting the default of the retirement saving rate. Employees hired before a set date were offered a choice with a default rate of 3%, while employees hired after that were offered the same choice with a default rate of 6%. The intended effect was an increase of those who chose the 6% contribution (from 24% to 49%) and a decrease of those selecting the 3% contribution (from 28% to 4%). It is tempting to consider this as evidence of the effectiveness of the default resetting in this particular context. However, such an interpretation neglects the effects the default resetting had at the individual level on other, non-default contribution rates. In particular, an analysis at the individual level showed that it reduced the frequency of higher contribution rates (the proportion of those selecting contribution rates over 7% fell from 41% to 30%) and increased the frequency of low contribution rates (the number of those selecting less than 3% increased from 5% to 15%). These effects are substantial: a revisit of the data shows that the mean contribution rate is actually lowered by the default resetting (from 6.88% to 6.31%). Consequently, the resetting did not just shift some people from a 3% to a 6% contribution rate. It also affected people with other contribution rates, so that (i) about the same number of people chose the new default and (ii) more people chose a lower contribution, yielding a mean reduction in contributions. Contrary to first impressions, the nudge produced a worse result because of the heterogeneity in the population. One can easily imagine similarly unwanted effects of the Save More Tomorrow (SMT) nudge if the heterogeneity of the underlying discounted evaluation is too large. In such a context, SMT would make some people save less, while other people would save more, without obtaining an optimum for anyone.

Some proponents of nudge policies do consider the possibility of such heterogeneity. *Asymmetric paternalism* (Camerer *et al.* 2003), for example, assumes that some members of a population may be fully rational, and hence not in need of a nudge that others require. Asymmetric paternalism refers to an intervention that 'creates large benefits to those who make errors, while imposing little or no harm on those who are fully rational' (Camerer *et al*. 2003: 1212). Consequently, it seeks to devise policies that affect only those 'in need' of a nudge. Asymmetric nudges are designed so that those whose behaviour is considered optimal or more than optimal remain unaffected and continue to implement their behaviour, even if deviant from the suggested option. Although a considerable improvement over the equal incompetence assumption, which runs into problems such as those described above, even asymmetric paternalism assumes that those who are subject to the bias are uniformly affected so that the same nudge can steer them toward more optimal choices. *Thus, the conclusion remains that to be effective, nudge interventions require sufficient homogeneity in the population's heuristic repertoire.*

4.5. Temporal effects

Nudges are designed to trigger the deployment of an existing heuristic to generate a certain behaviour, without trying to convince and sometimes without explicitly informing the agent about the goal of the intervention. That, of course, is the point of nudging: rather than convincing the agent that it would be in her best interest to save more or exercise more now, the nudge circumvents this difficult task by exploiting known tendencies or heuristics that reliably predict behaviours given certain environmental stimuli. The problem with this approach is that because the agent is not actively participating, she might engage in additional behaviour that counteracts or compensates the desired behavioural change.

Imagine, for example, a company policy that aims at making people exercise more, and thus subscribes all employees for 10 monthly gym visits by default. Assume that the policy makes a difference: once it is implemented, people spend more time in the company gym than before. However, spending more time in the gym is not the ultimate goal of the policy: making people exercise more is. It is possible that some people go more often to the company gym, but keep track of their overall exercise time and cut back proportionally on their previous exercise time outside the company gym. Whereas adaptation to boosts implies that a new and effective heuristic is learned and can be applied successfully, adaptations to a nudge are less predictable because the nudged are not consciously aware of the intervention and might be pursuing a different goal than the one intended by the policymaker. While the repeated and successful application of a heuristic acquired via a boost is likely to increase the agent's competence, the effect of a nudge might not persist over time because the repeated implementation adversely affects the stability of the triggering relationship between environmental stimuli and the agent's heuristics. *Thus, before designing and implementing a nudge, potential temporal effects, that is, reactions to a nudge over time, must be carefully analysed*.

Of course, over time the wearing-off effect might be counteracted by learning effects. For example, a study of energy conservation nudges shows that the effects of a one-time intervention wear off quickly. However, repeated exposure may 'gradually change [people's] capital stock of habits or physical technologies' so that 'after two years, … effects decay at only 10 to 20 percent per year' (Allcott and Rogers 2014: 3034). In this case, we suspect that a repeated nudge may come close to a boost in that it ends up slowly changing people's repertoire of strategies and/or resources for action.

Similar evidence concerning the malleability of people's heuristic strategies comes from research into default setting, which shows that after receiving a good default, people are more likely to accept defaults in the future, even if doing so results in worse outcomes than going against the default (de Haan and Linde 2016). The reverse effect is also possible, however. Let's consider the gym-exercise example again. One likely effect of increased gym attendance is that with time, employees will build preferences. That is, they will discover whether they like going to the gym and whether they feel the perceived benefit is worth their while. For people with clear-cut preferences, 'default rules are less likely to stick' (Sunstein 2016: 118). In that case, an initial increase in gym attendance due to the default might be followed by a drop-off, as only those who actually like going to the gym will stick with it.

Both types of effects – immediate compensations for nudges and changes in habits or preferences over time – might impact the effectiveness of nudge policies, as they modify the desired behaviour in hard-to-predict ways. More specifically, the very intervention that caused the behavioural change in the first place may cause its increase or degradation due to adaptations over time. Although such effects cannot be completely precluded for boosts, they are less likely to occur because the application of a newly learned cognitive heuristic is goal-directed by definition.[4] If, for instance, people apply Temptation Bundling (TB), they do so because they aim to increase their exercise regime. Consequently, they have little reason to compensate for more gym visits afterwards, unless they deliberately decide to do so. Nor should one expect the effect to wear off, because more use of the heuristic means more training in its application, and, as with most training, a more effective application as a consequence.

4.6. Selection criteria

To be effective, boosts require not only that agents learn a new heuristic, but also that they are able to recognize when to apply it under at least some range of variation in environmental conditions. More precisely, in order for a boost to reliably change behaviour in the long run, two conditions must be satisfied. The first is that the agents should be able to recognize the conditions for the effective selection of the new heuristic. The second is that the agents should continue to select the newly learned heuristic rather than fall back on previous heuristics. We focus on the former condition here and turn to the latter in the next subsection.

Boosts, such as Simple Rules of Thumb (SRT) and Temptation Bundling (TB), increase people's competence to flexibly apply heuristics in specific environments. As such, a heuristic 'may appear simplistic, but their underlying intelligence lies in selecting the right rule of thumb for the right situation' (Gigerenzer 2007: 49). To solve this so-called *strategy*

[4] The possibility that agents might use the heuristic in unexpected ways, which may undermine the effect that the newly taught heuristic was designed to achieve cannot be excluded a priori. It is however a far-fetched possibility. We thank an anonymous referee for pointing this out.

selection problem, people must be sensitive to their applicability criteria, which are hidden in the structure of cues in the choice environment. Take, for instance, the *recognition heuristic.* If you recognize the name of one city, football club or stock on a list but not the others, then this option is likely to rank highest in value and choosing it will likely be successful. However, the recognition heuristic breaks down if you recognize more than one option or if the reason why you recognized a particular option is no longer valid (e.g. the population size of Detroit decreased after the crisis of the automobile industry; the competitiveness of a football club decreased because its best players left, etc.).

Similarly, a simple lexicographic decision rule examines cues in a ranked order one at a time and makes a binary decision based on the first cue that discriminates between two options. Such a simple rule has been shown to perform at least as well as a linear decision algorithm that integrates all informational cues, but only if the environment is such that the predictive power of the highest-ranked cue indeed dominates all others and cannot be outweighed by the remaining cues (Şimşek 2013). Which cue structure is required for the effective application of a specific rule is an empirical question. The Fast and Frugal literature has shown that there are environments for which simple rules are a good match, but not all environments have a cue structure that can be matched successfully by some simple rule (e.g. Gigerenzer *et al*. 1999; Şimşek 2013). There might be environments in which decisions can be satisfactorily made only with computationally more intense strategies. *In sum, when the effectiveness of a boost depends on the cue structure of the environment, people should be taught not only the heuristic but also the right selection rule.*

4.7. Individual/environmental resources

What ensures that the agents will continue to select the newly learned heuristic rather than fall back on previous heuristics? Even when the agents are capable of recognizing the conditions for the application of the new heuristic, this in itself does not ensure that they will continue to select it. Other features of the decision environment, such as time, money, will power or value incongruence, might get in the way, especially in the long run. Consider for example the Temptation Bundling (TB) intervention. If personally relevant temptations cannot be effectively matched in terms of timing (e.g. I must go to the gym *and* want to go bungee jumping), financial means (e.g. I have money only to go to the gym *or* to go bungee jumping), or values (e.g. working out is healthy *and* eating cake is pleasant), it might be hard to stick with the new behaviour in the long run. In general, if resources such as time or money become scarce, it is relatively easy to fall back on pre-existing habits and behaviours. Similar problems occur if the new behaviours are not congruent with people's value systems. Even if in the beginning the tension is resolved in favour of the new behaviour, there is no guarantee that this will be maintained in the long run. *Thus, whereas nudges change environments to match people's heuristics by definition, when designing and implementing boostable heuristics, environmental conditions such as the particular value system of the targeted population and the (temporal and financial) resources available to them must be well analysed before a boost can be reliably implemented.*

5. CONCLUSION

We have distinguished boosts from nudges on the basis of the different mechanisms through which they operate. Our simplified models of these mechanisms consist of three main components: the environment, the agent's heuristic repertoire and her behaviour. In these models, nudges and boosts mainly differ with respect to the point of intervention: whether it is the environment (nudges) or the agent's heuristic repertoire (boosts). On the basis of these simplified models of the mechanisms, we have identified the contextual conditions that must be satisfied for nudges and boosts to be effectively applied.

In particular, we showed that, at the intervention stage, the effectiveness of nudges is mainly influenced by how reliably the designed changes in environmental stimuli trigger the intended behaviour (*trigger stability*). At the implementation stage, effective nudge interventions must consider the variance in strategies used across the targeted population as well as differences in levels of competence (*homogeneity of heuristic repertoire*). Similarly, we showed that the success of nudges is sensitive to potential *temporal effects*, such as when people adapt to nudges over time. For boosts, what mostly matters at the intervention stage are the *agents' motivation* to learn and implement a boost as well as its *teachability*. We also showed that for boosts to be implemented reliably, people must solve the *selection problem* by knowing when to apply a heuristic and have adequate *resources* to be able to apply it.

The success criteria we identified for boosts and nudges are not mutually exclusive. That is, there may be situations in which both interventions are effective or neither of them is. However, we believe that the results of our analysis can help to identify likely obstacles that a particular context of application might pose for the working of either mechanism, and hence serve as a planning tool for assessing and evaluating the likely effectiveness of a behavioural policy in a given context.

ACKNOWLEDGEMENTS

We would like to thank the organizers of the 13th Blankensee-Colloquium 'Nudge versus Educate: Comparing Two Approaches to Policy in Terms of their Modeling Strategies' as well as the Wissenschaftskolleg zu Berlin.

Their support provided the basis for the work presented in this article. Caterina Marchionni would like to acknowledge the financial support of the Academy of Finland.

REFERENCES

Allcott, H. and T. Rogers. 2014. The short-run and long-run effects of behavioral interventions: experimental evidence from energy conservation. *American Economic Review* 104: 3003–3037.

Bhargava, S. and G. Loewenstein. 2015. Behavioral economics and public policy: beyond nudging. *American Economic Review* 105: 396–401.

Benartzi, S. and R. H. Thaler. 2013. Behavioral economics and the retirement savings crisis. *Science* 339 (6124): 1152–1153. doi: 10.1126/science.1231320.

Beshears, J., J. J. Choi, D. Laibson and B. C. Madrian. 2009. The importance of default options for retirement saving outcomes: Evidence from the United States. *Social Security Policy in a Changing Environment*, ed. J. R. Brown, J. B. Liebman and D. A. Wise, 167–195. Chicago, IL: University of Chicago Press.

Bond, M. 2009. Risk school. *Nature* 461: 1189–1192.

Bovens, L. 2009. The ethics of nudge. In *Preference Change: Approaches from Philosophy, Economics and Psychology*, ed. T. Grüne-Yanoff and S. O. Hansson, 207–219. New York, NY: Springer

Brehm, S. and J. Brehm. 1981. *Psychological Reactance: A Theory of Freedom and Control*. London: Academic 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.

Campbell, M. C. 2007. 'Says Who?!' How the source of price information and affect influence perceived price (un)fairness. *Journal of Marketing Research* 44: 261–271.

Cartwright, N. 2010. Presidential address: Will this policy work for you? Predicting effectiveness better: how philosophy helps. *Philosophy of Science* 79: 973–989.

Cartwright, N. and J. Hardie. 2012. *Evidence-based Policy: A Practical Guide to Doing it Better*. Oxford: Oxford University Press.

Chetty, R. 2015. Behavioral economics and public policy: a pragmatic perspective. *American Economic Review* 105: 1–33.

Chow, S. J. 2015. Many meanings of 'heuristic'. *British Journal for the Philosophy of Science* 66: 977–1016.

Clarke, B., D. Gillies, P. Illari, F. Russo and J. Williamson. 2014. Mechanisms and the evidence hierarchy. *Topoi* 33: 339–360.

Craver, C. and J. Tabery. 2016. Mechanisms in science. In *Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. <https://plato.stanford.edu/archives/win2016/entries/science-mechanisms/>.

de Haan, T. and J. Linde. 2016. 'Good nudge lullaby': choice architecture and default bias reinforcement. *Econ J*. doi: 10.1111/ecoj.12440.

Drexler, A., G. Fischer and A. Schoar. 2014. Keeping it simple: financial literacy and rules of thumb. *American Economic Journal: Applied Economics* 6: 1–31.

Gigerenzer, G. 2007. *Gut Feelings: The Intelligence of the Unconscious*. New York, NY: Viking.

Gigerenzer, G., P. Todd and the ABC Research Group. 1999. *Simple Heuristics that Make us Smart*. Oxford: Oxford University Press.

Glennan, S. 2016. Mechanisms and mechanical philosophy. In *The Oxford Handbook of Philosophy of Science*, ed. P. Humphreys. Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780199368815.013.39.

Grüne-Yanoff, T. 2016. Why behavioural policy needs mechanistic evidence. *Economics and Philosophy* 32: 463–483.

Grüne-Yanoff, T. and R. Hertwig. 2016. Nudge versus boost: how coherent are policy and theory? *Minds and Machines* 26: 149–183.

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.

Haselton, M. G., D. Nettle and P. W. Andrews. 2005. The evolution of cognitive bias. In *The Handbook of Evolutionary Psychology*, ed. D. M. Buss, 724–746. Hoboken, NJ: John Wiley & Sons.

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.

Hertwig, R., G. Barron, E.U. Weber and I. Erev. 2004. Decisions from experience and the effect of rare events in risky choice. *Psychological Science* 15: 534–539.

Hertwig, R. and T. Grüne-Yanoff. 2017. Nudging and boosting: steering or empowering good decisions. *Perspectives on Psychological Science* 12: 973–986.

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D. and A. Tversky. 1996. On the reality of cognitive illusions: a reply to Gigerenzer's critique. *Psychological Review* 103: 582–591.

Kahneman, D., P. Slovic and A. Tversky. 1982. *Judgment under Uncertainty: Heuristics and Biases* (1st edn). New York, NY: Cambridge University Press.

Klein, E. 2011. Health care's brave new world of compulsory wellness. Reuters View, 12 October 2011. <http://www.bloombergview.com/articles/2011-10-13/health-care-s-brave-new-world-of-compulsory-wellness-ezra-klein>.

Loewenstein, G. and D. Prelec. 1992. Anomalies in intertemporal choice: evidence and an interpretation. *Quarterly Journal of Economics* 107: 573–597.

Ludwig, J., J. R. Kling and S. Mullainathan. 2011. Mechanism experiments and policy evaluations. *Journal of Economic Perspectives* 25: 17–38.

Marchionni, C. 2017. Mechanisms in economics. In *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, ed. S. Glennan and P. Illari, 423–434. London: Routledge.

Mega, L. F., G. Gigerenzer and K. G. Volz. 2015. Do intuitive and deliberate judgments rely on two distinct neural systems? A case study in face processing. *Frontiers in Human Neuroscience* 9: 1–15.

Milkman, K. L., J. A. Minson and K. G. Volpp. 2013. Holding the hunger games hostage at the gym: an evaluation of temptation bundling. *Management Science* 60: 283–299.

Mitchell, G. 2002. Why law and economics' perfect rationality should not be traded for behavioral law and economics' equal incompetence. *Geo. LJ* 91: 67.

Mongin, P. and M. Cozic. 2017. Rethinking nudge: not one but three concepts. *Behavioural Public Policy*, in press.

OECD. 2017. *Behavioural Insights and Public Policy: Lessons from Around the World*. Paris: OECD Publishing.

Oliver, A. 2013. *Behavioural Public Policy*. Cambridge: Cambridge University Press.

Pachur, T. and R. Hertwig. 2006. On the psychology of the recognition heuristic: retrieval primacy as a key determinant of its use. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31: 983–1002.

Polonioli, A. 2016. Reconsidering the normative argument from bounded rationality. *Theory and Psychology* 26: 287–303.

Rebonato, R. 2012. *Taking Liberties: A Critical Examination of Libertarian Paternalism*. New York, NY: Palgrave Macmillan.

Rieskamp, J. and P. E. Otto. 2006. SSL: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General* 135: 207–236.

Sampson, R. J., C. Winship and C. Knight. 2013. Translating causal claims. Principles and strategies for policy-relevant criminology. *Criminology and Public Policy* 12: 587–616.

Schubert, C. 2017. Exploring the (behavioural) political economy of nudging. *Journal of Institutional Economics* 13: 499–522.

Shafir, E. 2013. *The Behavioral Foundations of Public Policy*. Princeton, NJ: Princeton University Press.

Şimşek, Ö. 2013. Linear decision rule as aspiration for simple decision heuristics. *Advances in Neural Information Processing Systems* 26: 2904–2912.

Soll, J. B., K. L. Milkman and J. W. Payne. 2015. A user's guide to debiasing. In *The Wiley Blackwell Handbook of Judgment and Decision Making*, ed. G. Keren and G. Wu, 924–951. Oxford: Wiley Blackwell.

Steel, D. 2008. *Across the Boundaries: Extrapolation in Biology and Social Science*. New York, NY: Oxford University Press.

Sunstein, C. 2016. *The Ethics of Influence: Government in the Age of Behavioral Science*. Cambridge: Cambridge University Press.

Thaler, R. and S. Benartzi. 2004. Save more tomorrow: using behavioural economics to increase employee savings. *Journal of Political Economy* 112: 164–187.

Thaler, R. H. and C. R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New York, NY: Penguin.

Tversky, A. and D. Kahneman. 1974. Judgment under uncertainty: heuristics and biases. *Sciences* 185: 1124–1131.

Tversky, A. and D. Kahneman. 1986. Rational choice and the framing of decisions. *Journal of Business* 59: S251–S278.

van der Linden, S. 2011. Speed dating and decision making: why less is more. *Scientific American – Mind Matters (Nature)* http://www.scientificamerican.com/article/speed-dating-decision-making-why-less-is-more/

Volz, K. G., L. J. Schooler, R. I. Schubotz, M. Raab, G. Gigerenzer and D. Y. Von Cramon. 2006. Why you think Milan is larger than Modena: neural correlates of the recognition heuristic. *Journal of Cognitive Neuroscience* 18: 1924–1936.

BIOGRAPHICAL INFORMATION

Till Grüne-Yanoff is Professor of Philosophy at the Royal Institute of Technology (KTH) in Stockholm. He is also associated with the Centre for Philosophy of Social Science (TINT) at the University of Helsinki and the Max Planck Institute of Human Development in Berlin. His research focuses on philosophy of science, decision theory, and the relation of science and policymaking.

Caterina Marchionni is a Researcher at the Centre for Philosophy of Social Science (TINT), Social and Moral Philosophy, University of Helsinki. She is currently an associate editor of the *Journal of Economic Methodology* and the chair of the International Network for Economic Method (INEM). Her work mainly focuses on models, explanation, evidence and interdisciplinarity in economics and the social sciences.

Markus A. Feufel is Assistant Professor and heads the Division of Ergonomics at the Department of Psychology and Ergonomics at Technische Universität Berlin, Germany. His current work focuses on understanding and improving medical decision-making, risk communication and the impact of digital technology on work and decision processes. His main goal is to identify the prerequisites for informed decision-making and to develop interventions that help meet these prerequisites in work environments.