2019

Nudges that Should Fail?

Avishalom Tor
Notre Dame Law School, ator@nd.edu

Behavioural Public Policy, Page 1 of 27
 doi:10.1017/bpp.2019.5

Nudges that should fail?

AVISHALOM TOR*

Notre Dame Law School, Notre Dame, IN, USA and University of Haifa Faculty of Law, Haifa, Israel

Abstract: Professor Sunstein (2017) discusses the possible causes for and policy implications of the failure of nudges, with special attention to defaults. Though he focuses on nudges that fail when they should succeed, Sunstein recognizes that some failures reveal that a nudge should not have been attempted to begin with. 'Nudges that fail', however, does not consider fully the relationship between the outcomes of nudging and their likely welfare effects, most notably neglecting the troubling case of nudges that succeed when they should fail. Hence, after clarifying the boundaries of legitimate nudging within a libertarian-paternalistic approach and noting the fourfold relationship between the efficacy of nudging and its normative desirability, this article evaluates more fully the case of failed nudges and examines the hitherto unaddressed problem of successful yet undesirable nudges. This analysis shows that the failure of nudging bears only limited diagnostic value, while the success of a nudge is even less indicative of its normative status. The article concludes with recommendations for policy-makers who wish to employ nudges that are not only efficacious, but also likely to advance the subjective well-being of the individuals they target.

Submitted 27 November 2018; revised 28 January 2019;
accepted 11 February 2019

Professor Sunstein's recent article in this journal entitled 'Nudges that fail' (2017) discusses some of the possible causes for and policy implications of the failure of nudges, with a special focus on the paradigmatic case of default rules. The article correctly notes that the failure of a nudge can result from a variety of causes that, in turn, bear different policy implications, from nudging more effectively, through avoiding the nudge altogether, to replacing the attempted nudging with 'harder' interventions. The present analysis continues this examination, assessing more fully the relationship between

* Correspondence to: Professor of Law and Director, Research Program on Law and Market Behavior, Notre Dame Law School (ND LAMB), 1100 Eck Hall of Law, Notre Dame, IN 46556, USA. Email: ator@nd.edu

the efficacy of nudging and its potential for achieving its stated goal of advancing individuals' subjective welfare or, more colloquially, of making people better off according to their own lights (Thaler & Sunstein, 2008; Tor, 2017).[1]

A two-by-two table (see Figure 1) illustrates the fourfold relationship between a nudge's success or failure and whether it should be attempted in a given case. This simple framework makes clear that – as its title clearly indicates – Sunstein's article considers only the right-hand column of the table: the cells involving nudges that fail. The top-right cell concerns situations in which nudging fails even though it is normatively desirable, because its success would have advanced individuals' subjective welfare. This type of failure reveals the need to fix the attempted nudge.

Thus, if the failed nudge was only technically deficient, it merely needs to be better designed. To illustrate, consider a company that finds the majority of its employees opting out of a default contribution rate of 10% of their pretax income (cf. Beshears *et al.*, 2010), a default that the company adopted to help increase the retirement savings of those employees who wish to do so. This nudge might have failed due to some technical deficiency in its design, such as the default contribution rate having been set too high. If this were the case, employees would have found a better-designed default (e.g., a 6% contribution rate) more attractive, with overall contribution levels increasing and only a few employees opting out and further lowering their contribution rates.

Sunstein also considers two situations in which a nudge's failure places it in our bottom-right cell. In the first case, individuals' subjective welfare can still be improved by some other intervention, although no nudge would prove adequate to the task. According to 'Nudges that fail', such inadequate nudges should be replaced by 'harder' rules that diminish actors' freedom of choice (e.g., by changing economic incentives) or even deprive them of it altogether (as when adopting mandates or bans). In the case of the failed 10% retirement contribution default, for instance, the argument might be that if no default rate could bring employees to save as much as they wish to save, some mandatory minimum contribution rate should be set instead.

1 Importantly, though the analysis here has implications for behavioral interventions based on deontological, social welfare or traditional paternalistic grounds, it focuses on assessing libertarian–paternalistic nudges (Sunstein & Thaler, 2003) without addressing the legitimacy of the former grounds for policy-making. To this end, and despite the limitations of such an approach, the article follows the practice of mainstream economics, law and economics and public policy research, which commonly equates subjective welfare with actual preferences (Hausman, 2012) instead of drawing on other theories of well-being (Zamir, 1998).

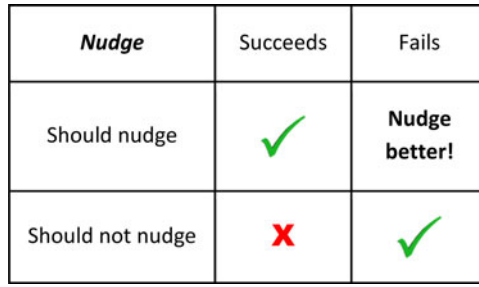

Figure 1. Nudge desirability and outcomes.

Inappropriate nudges – cases in which the failure of nudging is due to policy-makers' erroneous assessment of individuals' preferences – also belong in the bottom-right cell of Figure 1. Nudges cannot advance subjective welfare when unbiased individuals make contrary choices based on their antecedent preferences.[2] For example, one field study that manipulated the accessibility of different types of bread in two Dutch supermarkets found no significant effect on customers' propensity to purchase more versus less healthful types of bread (de Wijk *et al.*, 2016). While the failure of this particular nudge might have reflected the weakness of the experimental manipulation (a mere reordering of the presentation of most bread types within their usual aisle), it could well have been due to the clear antecedent preferences of consumers, who frequently shop for bread and are familiar with the available varieties. The choices of those who prefer and routinely purchase white or wheat bread, for instance, are unlikely to be affected by a minute change in their relative accessibility. Plausibly, the failure of this manipulation may simply reveal that shoppers cannot be made subjectively better off by a nudge involving the in-store location of different bread types. If this indeed were the case, policy-makers seeking to promote individuals' subjective welfare should have avoided further nudging (or any alternative intervention) attempts.[3]

However, while 'Nudges that fail' begins the important task of identifying different reasons for nudge failure, it stops short of explaining how one might go about determining the cause of a given failure. Yet, without such a determination, policy-makers would not be able to decide whether the

2 While interventions that transform people's preferences sometimes may be justified (Lehwinson-Zamir, 2015), they are largely inapposite with an approach that aims to advance the satisfaction of actual antecedent preferences (Sunstein, 2018a), though a full discussion of this question is outside the scope of the present analysis.

3 Of course, those wishing to promote goals such as social welfare or even individuals' objective welfare might still wish to replace the nudge with harder regulatory tools (e.g., Bubb & Pildes, 2014).

failure of a nudge should be endorsed or, instead, remedied. They would also be unable to determine whether a failed nudge that requires remediation should be replaced by a better nudge or by a more assertive policy instrument.

Even more problematically, Sunstein's focus on failed nudges diverts attention from nudges that belong on the left-hand side of Figure 1 – namely, nudges that succeed. The effective and desirable nudges of the top-left cell require no further discussion. But the case is quite different with respect to those nudges in the bottom-left cell that succeed when they should have failed. This category thus concerns undesirable nudging that should never have taken place to begin with, yet was successfully implemented.

The present article responds to these lacunae, examining how to determine whether the failure of a nudge is diagnostic of its undesirability; whether a failed yet-desirable nudge should be replaced by a better nudge or by 'harder' policy instruments beyond nudging; and, importantly, whether a successful nudge in fact should have failed. This inquiry shows that the failure of a nudge often holds only limited diagnostic value. Moreover, the apparent success of a nudge typically is even less informative than its failure. After explaining the substantial challenges involved in correctly categorizing both failed and successful nudges, this article concludes by outlining concrete recommendations for policy-makers wishing to employ efficacious nudges that may also advance individuals' subjective welfare.

Goal and tools – the why and how of true nudging[4]

Recent years have seen a dramatic rise in the study and implementation of behaviorally informed public policies (Shafir, 2013). One important catalyst for this development was the 2008 publication of Thaler and Sunstein's *Nudge: Improving Decisions about Health, Wealth, and Happiness* (hereafter referred to as *Nudge*), which brought to international and popular attention the developing academic discourse in this area. Indeed, *Nudge* has been so successful that much of the wide-ranging work by scholars and policy-makers over the decade since the book's publication has been referring to any and all behavioral public policies as 'nudges' (e.g., Halpern, 2015; Sibony & Alemanno, 2016).

However, the popularity of the nudge usage obscures an important aspect of Thaler and Sunstein's original definition of the term. As originally conceived, nudges are one important class of behaviorally informed policies, distinct

4 This section draws on Tor (2016, 2017). Compare also the discussion of the ends and means of behaviorally informed policies in Zamir and Teichman (2018).

from other such interventions in combining the important yet limited policy goal of improving individuals' subjective welfare with a specific set of liberty-preserving behavioral policy tools (Thaler & Sunstein, 2008; Tor, 2016). This unique 'libertarian-paternalistic' combination of goal and tools is the foundation of nudges' claim for superiority to behavioral policies that aim to advance other goals, such as social welfare, as well as to 'harder' paternalistic interventions (Sunstein & Thaler, 2003).[5] To appreciate this distinction and its relevance here, let us consider briefly the goal of nudging and its tools.

Nudging is self-consciously paternalistic, proclaiming the goal of making the individuals it targets better off. Yet Thaler and Sunstein make it clear that not all traditionally paternalistic interventions count as nudges, only those aiming at making people 'better off, as judged by themselves' (Thaler & Sunstein, 2008; Sunstein, 2018a). Thus, in contrast with traditional paternalism, which aims to improve individuals' welfare as the paternalist judges the matter, nudging is more constrained. A traditionally paternalistic policy, for instance, may discourage individuals from smoking marijuana or driving a car without wearing a seatbelt because policy-makers believe these activities are harmful to those engaging in them. Such paternalistic polices are not based on the subjective preferences of their targets, which may or may not agree with policy-makers' views. Furthermore, an individual wishing to smoke marijuana or intending to drive without a seatbelt may engage in these acts either based on a full understanding of the risks involved or due to some misunderstanding of said risks on her part. In either case, however, the subjective beliefs or preferences of the targeted individuals have only a limited bearing on the policy's desirability from the perspective of the paternalist policy-makers, who believe they know better.[6]

Nudges are different, involving only a paternalism of means rather than the traditional paternalism of ends (Sunstein, 2013). But making people subjectively better off as these people judge the matter requires policy-makers to be attuned to individuals' beliefs and preferences. This goal still leaves room for paternalistic interventions, because people may wish to act one way – say, to avoid smoking marijuana – yet find it difficult to implement their preferred course of behavior. Individuals may also neglect to take actions they would have liked to take, such as failing to wear a seatbelt while driving, due to lapses of attention rather than following conscious decisions. Additionally, people can benefit from paternalistic interventions when they hold erroneous

5 For a related but distinct approach, see Camerer *et al.* (2003).

6 Some traditional paternalists may consider individuals' subjective beliefs and preferences as one of many factors that affect these individuals' overall well-being, but neither as the main factor in their calculus nor as a determinative one.

beliefs that lead them to intentional acts that in fact conflict with their own preferences. For example, some individuals who consciously refrain from wearing a seatbelt while driving may do so only because they underestimate the risks involved. Had they been fully cognizant of these risks, these actors would have worn a seatbelt, so an intervention that leads them to do so may improve their subjective welfare.

Thaler and Sunstein (2008) believe that the goal of advancing individuals' subjective welfare is best achieved by employing a limited set of nudging tools "that alter people's behavior in a predictable way without forbidding any option or significantly changing their economic incentives" (Thaler & Sunstein, 2008, p. 6). Such choice-preserving tools are only a subset of the many policy instruments available even to strict means-paternalists, who can also employ, in principle, any of the more coercive policy tools that traditional ends-paternalism often adopts. To illustrate, a means-paternalist who believes that individuals fail to wear seatbelts only because they underestimate the risks involved can mandate that drivers wear seatbelts and penalize drivers who fail to comply.

Determining which tools are best suited to improving subjective welfare in any particular case would have been simple if means-paternalists were always and perfectly able to identify people's beliefs and preferences. Possessing full information and facing no risk of error, such policy-makers could have employed the most efficient tools available to them, accounting for the efficacy of alternative policies and their attendant costs. Efficacy considerations, for instance, often favor more coercive policies such as mandates or bans accompanied by sanctions for violation, which can be quite effective in changing behavior (cf. Zamir, 1998). At the same time, the implementation of these coercive policies often involves substantial enforcement costs (Sibony & Alemanno, 2016). Choice-preserving policies, on the other hand, may be less efficacious, since by design they allow individuals to act contrary to those true preferences that we assume the means-paternalist policy-makers to have identified. At the same time, such policies may be attractive due to their relatively low implementation costs, since enforcement is unnecessary when choice is preserved.[7]

In reality, however, means-paternalists may err when grappling with the daunting challenge of determining individuals' subjective beliefs and preferences (Rizzo & Whitman, 2009; Sunstein, 2014), a task that ends-paternalists

7 The implementation of choice-preserving policies may entail other substantial costs, such as when information needs to be disseminated widely, but the question of implementation costs is not germane to the present analysis.

are largely exempt from. Because policy-makers do not have direct access to this subjective information, they must infer it from observed conduct or statements, policy-relevant research and similar sources. Such inferences are inevitably fallible, however, and the adoption of policies seeking to advance the wrong preferences can impose substantial additional social costs. Means-paternalists must therefore also account for error costs, in addition to efficacy and implementation costs, when selecting their policy tools.

Once error costs are considered, though, nudges seem to hold the upper hand relative to 'hard' policies. Due to their more coercive nature, the latter tools offer only limited feedback, if any, about policy-makers' successes in identifying individuals' preferences. In the rare, extreme cases in which even 'hard' policies fail to change the behavior of their targets, individuals' resistance likely indicates a dramatic discrepancy between their own preferences and policy-makers' beliefs about these preferences. More commonly, however, policies involving greater coercion will successfully change behavior, whether through legal mandates or bans or via other economic incentives, while offering little insight regarding people's antecedent preferences. The changes wrought by such policies may simply reflect individuals' constrained choices or their responses to altered incentives rather than their subjective views of what makes them better off. In these common cases, therefore, the information loss resulting from the employment of more coercive policy instruments entails the risk of substantial error costs.

In contrast, nudges appear more informative and less costly for means-paternalists who are concerned about the possibility of error in ascertaining subjective beliefs and preferences. These choice-preserving policies try to steer people toward their desired goals while leaving them free to go their own contrary way. Consequently, the behavior of individuals who accept a policy intervention that they are free to reject would seem to suggest the nudge is compatible with their subjective beliefs and preferences. The failure of a nudge, on the other hand, alerts policy-makers to the possibility they erred in their judgment of their targets' subjective perspective, thus allowing means-paternalists to correct their errors and avoid their long-term costs (Sunstein, 2017).

The potential error-cost advantage of nudges compared to coercive interventions is even more significant once the heterogeneity of preferences is accounted for. Different people hold different preferences in all policy-relevant domains of human behavior, from health and finances to labor, leisure and more. This heterogeneity makes the task of means-paternalists particularly difficult, since no single policy can advance the subjective well-being of all of its targets.

Furthermore, the selection of specific policies faces an even greater challenge in the many common cases that require decision-makers to trade off different

values against one another. Most people, for example, prefer larger to smaller retirement savings. In the absence of tradeoffs between an increased rate of saving for retirement and any other consideration, means-paternalists may implement policies that help people achieve their desired goal with reasonable confidence. Yet in reality, individuals who save more of their current income for retirement inevitably reduce their disposable resources for present consumption, and the relative weight that different people give to savings versus consumption is bound to vary. Thus, a 'hard' measure, such as a mandatory minimum retirement contribution rate, will increase contribution rates but reduce the disposable income of employees who were saving less before its implementation. The effect of the mandate on the employees' subjective well-being will depend on their preferences, however. Those who wished to save more but failed to do so prior to the mandate will be better off, while their peers who preferred having more disposable income at present will be subjectively worse off. On the other hand, a choice-preserving default contribution would allow the employees who wish to save less to do just that, even while helping those who wish to save more increase their retirement savings.

The preceding analysis helps explains *Nudge*'s insistence on employing choice-preserving tools when aiming to improve individual subjective welfare. When practical, the unique libertarian-paternalistic combination of goal and tools appears particularly attractive, even apart from any libertarian commitment to individuals' freedom of choice. Policies that seek to make individuals better off as they see the matter and use only choice-preserving nudges should be able to avoid the potentially costly imposition of policy-makers' sometimes mistaken views of what makes other people better off.

Nudges that fail

As noted, 'Nudges that fail' addresses three categories of nudge failure that we may designate technically deficient nudges, inadequate nudges, and inappropriate nudges. Naturally, the designation of a failed nudge as belonging to a particular category has significant policy implications. We saw that technically deficient nudges simply require fixing. The actual task of identifying the precise contours of an effective, improved nudge may be demanding, but policy-makers need not be concerned about the normative desirability of nudges that properly belong to this first category.

The clear prescription for inadequate nudges, on the other hand, is to replace them with more coercive policies. This second category involves interventions that even committed means-paternalists should find normatively justified because their success advances subjective welfare regardless of the specific tools they employ. Importantly, moreover, a correctly designated, inadequate

nudge by definition raises no error-cost concerns; although nudging proved insufficient to the task, achieving the behavioral change sought by the failed nudge will make its targets better off. This absence of error costs, in turn, combines with the superior efficacy of harder policy measures to make them a natural substitute for the failed, softer nudge.

The third category – that of inappropriate nudges – bears a dramatically different prescription. These are cases in which the failure of nudging is diagnostic, informing us that the nudge should never have been attempted. Notably, unlike the preceding two categories, inappropriate nudges take place only when policy-makers' initial assessment of individuals' beliefs or preferences was erroneous. From the perspective of means-paternalists, the failure of inappropriate nudges is therefore fortunate. It is a case in which 'freedom worked' (Sunstein, 2017), a case that vindicates the softer touch of the nudging tools that enable people to make contrary choices without facing legal or substantial economic costs.

Yet the otherwise-insightful analysis of 'Nudges that fail' stops short of acknowledging the critical challenge of appropriately categorizing failed nudges. In other words, even means-paternalists who fully endorse the distinction proposed here among technically deficient, inadequate, and inappropriate nudges are left with no clear guidance on how to determine which category a given failed nudges properly belongs to. The article does seem to suggest in passing a number of solutions to this challenge, including relying on policy-makers' judgments of what makes other people better off and ascertaining whether the nudge failed due to its targets' strong antecedent preferences or because of some judgmental bias on their part. However, further reflection reveals the problematic nature of all of these solutions.

The suggestion that even following the failure of a nudge we continue to rely on policy-makers' judgments of individuals' preferences, particularly for low-cost nudges or nudges that address significant problems, is troublesome. Concluding that a failure is not diagnostic simply because policy-makers believe they know better altogether negates the very error-cost advantage of nudging compared to harder interventions. Indeed, once adopted, such a conclusion risks turning means-paternalists into the same ends-paternalists from whom they seek to distinguish themselves. It is still possible, of course, that these policy-makers have correctly identified individuals' preferences, but the intuitions of means-paternalists offer a thin reed to hold on to in the face of a nudge that failed due to something other than its technical deficiency.

Furthermore, a continued reliance on policy-makers' judgment of what people really want would still fail to discriminate between technically deficient nudges and inadequate nudges. Perhaps the implied suggestion is that confident means-paternalists should follow a failed nudge with further

nudging attempts and switch to a harder intervention only if these further nudges continue to fail. Such a practice is unattractive, however, since it involves continued lower-efficacy, soft interventions in situations that render their error-cost advantage irrelevant because policy-makers already believe they have nothing new to learn about their targets' subjective preferences. Instead, assuming their nudge already was reasonably designed, those who are confident in their subjective welfare judgments will be better served by adopting more coercive policies immediately on the heels of the failed nudge.

The difficulty with continuing to rely on policy-makers' own judgments following the failure of nudging is obvious, however. Means-paternalists who are genuinely concerned about the risk of misjudging their targets' preferences will be unlikely to abandon their concern in the face of potential evidence that a failed nudge so usefully offers for precisely such a mistake. Their counterparts who believe that they are not prone to error, on the other hand, need not await the failure of a nudge to adopt more coercive policies and will see little benefit in the choice-preserving tools of nudging to begin with (Conly, 2013).

A second, more attractive approach to distinguishing among the three failed nudging categories is to determine whether a given failure reflects individuals' strong antecedent preferences or is rather due to their lingering judgment errors. 'Nudges that fail' suggests the relevance of this approach when it discusses both examples of nudges that apparently failed due to robust antecedent preferences and other cases in which a failure may be attributed to individuals' naturally arising or induced biases. The former examples are the poster children of a diagnostic failure. The subjective well-being of unbiased individuals who refuse to be nudged for the right reasons cannot be improved through further intervention, and means-paternalists will happily leave these people to their own devices.

In contrast, nudges that fail at least in part due to people's biased judgments present a more complicated picture. Such policies in fact may turn out to belong to any of the three categories of technically deficient, inadequate or inappropriate nudges. To illustrate the challenge of properly categorizing failures involving biased judgments, let us return to the case of a company that wishes to help its employees increase their retirement saving by using a default retirement contribution rate (e.g., 6% of salary). Assume that most employees opt out of the default and choose instead to save a smaller percentage of their employment income, but the company is unsure as to the precise reasons for the failure of the nudge. The company also finds evidence that many of its employees manifest one or more biases that lead them either to overweight present versus far-future consumption in retirement or to overestimate their likely cumulative savings at retirement as a function of their current contribution rate. For example, these employees may be present-biased,

hyperbolic discounters or, alternatively, just overoptimistic about the long-term performance of their retirement investment portfolio.

Notably, in either case, it does not suffice for the company to determine that its employees are biased. As long as the magnitude of the employees' errors, their effect on the decision to opt out from the retirement contribution default, and the impact of employee preferences for present versus retirement consumption are unknown, the company will not be able properly to categorize the nudge's failure. It may be, for example, that the failure is wholly or mostly due to the employees' biased judgments, in which case the company would have to decide whether to attempt to debias these judgments so that the nudge will work, essentially treating the failure as a technically deficient nudge. Alternatively, the company might accept the bias as given and replace what it views as an inadequate nudge with a mandatory contribution rate. Yet the employees' behavior could also be significantly driven by strong antecedent preferences for present consumption, regardless of their judgment bias. In the latter case, however, even fully debiased employees will continue to opt out of the default, and so the company has attempted an inappropriate nudge.

At any rate, the broader and more significant point is that even reliable evidence that a bias contributed to the nudge's failure bears unclear ramifications without additional information. Policy-makers who can further determine the relative impact of biased judgments on the failure of a nudge compared to the role played by individuals' antecedent preferences will have a clearer path forward. Yet such information is difficult to come by and unlikely to be available for most failed nudges. Rather, a more likely finding that biased judgments made some contribution to the failure of a nudge would not significantly ameliorate the challenge of correctly categorizing the failure in order to determine the appropriate follow-up policy response.

Finally, the already difficult task of categorizing failed nudges is further complicated by the limited correlation between the strength of individuals' antecedent preferences and their resilience to nudging. Sunstein (2017) makes the intuitively appealing argument that the strength of individuals' antecedent preferences should manifest in their resilience to nudging, at least in the absence of judgmental biases. On this account, people with strong preferences find a contrary nudge more costly (whether consciously or intuitively) and therefore resist it, while those who possess only weak antecedent preferences, or none at all, succumb to the same nudge. If this were the case, all properly designed nudges directed at unbiased targets would have had the salubrious effect of advancing the subjective welfare of individuals not possessing contrary preferences while allowing those holding contrary preferences to go their own way with minimal impediments.

Yet behavioral research clearly contradicts this intuition, showing that there is substantial heterogeneity in individuals' tendencies to approximate the assumptions of rationality or deviate from them (Stanovich *et al.*, 2008). The empirical evidence finds, moreover, that intra-individual correlations in the manifestation of different behavioral phenomena are quite low (Tor, 2014, 2015). Hence, for example, an overoptimistic decision-maker may be impervious to framing effects. Decision-makers' susceptibility to nudging generally and defaults specifically will vary greatly, depending *inter alia* on the individual and the specific nudging technique employed. The joint effects of heterogeneity in both domains of preferences and rationality thus mean that individuals' reactions to nudging will not reliably reflect the strength of their antecedent preferences with any reliability. In the illustration of a company's retirement contribution policy, employees who are largely unaffected by defaults may opt out despite holding very weak antecedent preferences for current consumption, contributing to the nudge's failure.[8]

All in all, the challenge facing any effort to categorize failed nudges properly is substantial, particularly once the limits of the information that policy-makers are likely to possess even in the best of cases are taken into account. It is the difficulty of properly categorizing failed nudges, therefore, rather than the more straightforward normative implications of properly categorized failures, that should be the primary concern of committed means-paternalists.

Nudges that should fail?

The preceding analysis highlighted the challenge of properly categorizing failed nudges, showing that neither policy-makers' beliefs about the subjective welfare of others nor their efforts to distinguish bias-driven from preference-wrought failures are likely to offer reliable guidance for post-failure policy selection. Another challenge that 'Nudges that fail' stops short of addressing – namely, the problem of determining the normative desirability of nudges that have *succeeded* – is even more troubling, however.

Why should means-paternalists be concerned about the desirability of successful nudges? The answer is simple: Successful nudges, just like failed ones, may turn out to be inappropriate. Unlike the other two categories of technically deficient and inadequate nudges, which apply only to failed nudges, the question of whether a nudge should have been attempted to begin with – namely, of

8 The opposite and normatively more troubling pattern, in which employees who are more susceptible to the effects of defaults are successfully nudged even when they hold contrary preferences, is discussed below.

potentially inappropriate nudges – applies to successful ones as well, and with greater force. To wit, the failure of a nudge at least alerts policy-makers to the possibility that the nudge was inappropriate, the difficulty of properly categorizing the failure notwithstanding. Yet the same cannot be said of successful nudges, which may succeed in changing the behavior of individuals who hold contrary antecedent preferences without leaving any indication of their inappropriateness.

Successful nudges can sometimes override the preferences of individuals whose pre-nudge judgments are biased in a direction that makes the nudge appear to promote their preferences when its actual effect is to the contrary. To illustrate, an employee who underestimates the benefits of compound interest for the long-term growth of her retirement savings may retain the high default contribution rate set by the company because she erroneously believes she must save more now to maintain her preferred level of retirement savings. In this case, by leading her to save too much and consume too little at present, the successful default contribution nudge diminishes the employee's subjective welfare instead of advancing it.

Although the concern that some previously biased individuals are susceptible to inappropriate nudges is real, its practical effects are likely to be limited. After all, those whose judgments are biased in a direction that makes them excessively likely to be nudged successfully will also be more likely to make choices that diminish their subjective welfare even in the absence of the nudge. The employee who underestimates the expected growth of their long-term savings, for example, will tend to save excessively regardless of the employer's adoption of any contribution default. Nonetheless, a more passive employee or one who is only modestly biased may end up saving excessively at present only when nudged in that direction, but not if they are left to their own devices.

The problem of bias-based susceptibility to mistaken nudging is more significant, however, for those individuals who are biased by the nudge itself. Specifically, some people may accede to nudges partly because they believe them to convey valid information about desirable behavior.[9] For instance, an employee facing a 6% default contribution may infer that this level of contribution is necessary to reach her retirement savings goals and therefore allow herself to be nudged. Yet if this employee's goals are achievable with a lower

9 Some nudges may also seem to convey information that changes people's preferences rather than only shape their beliefs, as discussed below. A nudge that appears to convey a prevailing social norm, for instance, may lead some individuals to change their preferences so that they better align with that norm (cf. Bar-Gill *et al.*, 2018).

contribution (e.g., 4%), the successful nudge will diminish her net present income and thus reduce rather than increase her subjective welfare.

Yet another problematic case involves inappropriate nudges that successfully change the behavior of individuals who hold contrary preferences not only prior to but even following the nudge. At first blush, such inappropriate success may seem implausible given the choice-preserving nature of nudges. Since the nudged are free to go their own way without bearing substantial economic costs, one might expect those whose preferences disagree with the nudge simply to disregard it. Indeed, the freedom of contrary choice feature of nudging is its hallmark advantage over traditional, more coercive policy tools such as mandates, bans or taxes (Sunstein, 2014, 2017). Moreover, empirical studies of nudging often show evidence of some nudge-resistant choices, as when individuals opt out of default arrangements (Beshears *et al.*, 2010; Willis, 2013).

A closer look reveals, however, that nudges can alter behavior and override individuals' antecedent preferences even while preserving their targets' nominal freedom of choice. To understand how this can happen, we should recall that many nudges succeed by employing behavioral tools that are only (or mostly) efficacious because of individuals' bounded rationality. This is the case with nudges that influence judgment processes, such as those that rely on anchoring or availability, but it also true for nudges that directly impact choice behavior. Nudges that shape choice by exploiting, for example, ordering or context effects, framing effects or loss aversion (for a more extensive list, see Sunstein, 2016) by definition affect only boundedly rational individuals. The hypothetical, perfectly rational actor – the imaginary breed that Thaler and Sunstein (2008) designate 'Econs' – would have made the same choices regardless of the order in which options were presented, their context, frame and so on. In fact, the behavioral research program that over the last half-century has documented most of the phenomena that today are advanced as potential reasons for nudging, as well as useful nudging tools, is largely based on a systematic study of how actual human judgment and decision-making differ from the implicit assumptions and explicit axioms of rational choice (Goldstein & Hogarth, 1997).

The same bounded rationality of real individuals, which opens the door to so many forms of efficacious nudging, also means that inappropriate nudges may succeed by transforming their targets' antecedent preferences or simply overriding them (Tor, 2016, 2017). For instance, some boundedly rational employees may be successfully nudged toward higher retirement contribution rates even when they hold antecedent preferences for more consumption and therefore a lower savings rate at present (cf. Johnson & Goldstein, 2013). In this case, the default contribution nudge may be effective because it changes the

preferences of these employees, who now truly wish to consume less of their income and save more for retirement. Such an outcome, in which a nudge transforms its targets' preferences, is inappropriate since nudging seeks to make people better off as they see the matter, rather than to change their views concerning what makes them better off.[10]

Alternatively, the default contribution may nudge some employees to increase their savings even though they retain their antecedent preference for higher present consumption and a concomitantly lower saving rate. This outcome is inappropriate for a different and obvious reason – that is, because the nudge indisputably reduces the subjective welfare of its targets. Such inappropriate nudging may be successful when people hold weak or imperfectly formed preferences or because some of the targeted individuals are more susceptible to the particular nudge's influence.

Though still troubling, the former case is perhaps of lesser concern because interventions that override weak or imperfectly formed preferences are less likely to cause substantial reductions in subjective welfare. A successfully nudged individual who only slightly prefers an alternative outcome, or who merely thinks she probably prefers that other outcome, may face a smaller gap between her subjective preference and her post-nudge behavior than one whose preferences significantly and clearly diverge from her post-nudge behavior. To illustrate, an employee who retains a 6% contribution default when she would have slightly preferred contributing only 5%, or thinks she probably would have preferred to contribute only 4%, is likely better off than one who is certain she would have preferred to contribute only 4%.

A nudge may succeed in changing the behavior of some individuals who retain their contrary antecedent preferences because different people are differently susceptible to behavioral interventions. Research already mentioned reveals much heterogeneity in the degree to which decision-makers approximate rational judgment and decision-making. In other words, individuals' bounded rationality manifests differently with respect to different behavioral phenomena, in different contexts and times and so on (Tor, 2014). Consequently, two different, unbiased individuals with similar antecedent preferences may

10 One might also argue that the nudge promotes these individuals' subjective welfare as measured by their actual preferences following the nudge (Zamir & Medina, 2010; Sunstein, 2018a). Yet, though a systematic analysis of the status of preferences *ex post* nudge versus *ex ante* nudge is outside the scope of the present article, note that counting interventions that change individuals' preferences as nudges vitiates much of the advantage of the subjective welfare standard in disciplining potential nudges, whatever other merits such preference-transforming interventions may possess (Lehwinson-Zamir, 2015). This approach also opens the door to public choice concerns regarding the potential employment of nudging to manipulate citizens' preferences (Glaeser, 2006) and related critiques (Rizzo & Whitman, 2009).

well respond differently to a nudge in a direction that is contrary to these preferences, one resisting the nudge even while the other succumbs to it. For this reason, the success of any given nudge in changing the behavior of its unbiased targets depends not only on the existence, direction and strength of their antecedent preferences, but also on their susceptibility to the particular nudge that means-paternalists decide to employ.

Therefore, the ultimate population of those who are successfully nudged in any given case is likely composed of three subgroups that policy-makers will find difficult to distinguish from one another: (1) those whose antecedent preferences are aligned with the nudge; (2) those holding weak or imperfectly formed contrary preferences; and (3) those who are more susceptible to the particular method of nudging and therefore change their behavior despite retaining their clear contrary antecedent preferences. The first of these groups involves cases of desirable, appropriate nudging; the second group includes those somewhat troubling, borderline instances; but the third group is the most problematic, since it is composed of cases of unquestionably inappropriate nudging.

Returning to our familiar retirement contribution default example, a company adopting the 6% default may find that most employees do not opt out and that overall employee retirement savings increase as a result. If post-nudge behavior were a reliable indicator of antecedent preferences, the company could have reasonably concluded that its nudging effort was both successful and appropriate. However, given the heterogeneity in its employees' susceptibility to the default rule nudge, the company does not know whether the subjective welfare of a significant portion of those who were successfully nudged perhaps was decreased, rather than increased.

All in all, the challenge posed by the need to identify correctly instances of inappropriate yet successful nudging is substantial. Successful nudges may be inappropriate because their success is due their targets' extant or nudge-induced bias. They may also be inappropriate because they have succeeded by transforming – or simply overriding – the preferences of the very individuals whose subjective welfare means-paternalists wish to promote. Much like in the case of nudges that fail, moreover, policy-makers will find it difficult to distinguish these varieties of inappropriate, successful nudges from their desirable, appropriate counterparts. However, unlike nudges that fail, whose failure at least reveals the need to assess their appropriateness, nudges that succeed offer no such signal. The very sought-after success of the latter provides neither a negative nor a positive indication of their desirability, leaving the conscientious means-paternalist wondering whether these successful nudges should have failed.

The way forward

Policy-makers wishing to promote individuals' subjective welfare face a conundrum. On the one hand, true nudging promises to be a superior means for helping make people 'better as they judge the matter' compared to traditional, more coercive regulatory instruments. On the other hand, even well-informed, conscientious means-paternalists cannot reliably distinguish failed nudges that should have succeeded and thus ought to be fixed from those undesirable ones whose failure is diagnostic. Moreover, the identification of successful nudges that should never have been attempted is an even greater challenge.

Nudge skeptics facing this conundrum may quickly conclude that faithful nudging is impossible, since one cannot reliably advance individuals' subjective welfare through choice-preserving behavioral interventions. Some of these skeptics will take this conclusion further to justify a rejection of both 'libertarian paternalism' and its tools, while others may determine that even means-paternalists should routinely employ traditional 'hard' regulation.

Yet there is a way forward even for committed subjective welfarists. Specifically, the following paragraphs briefly describe two related, more measured responses to the challenge involved in determining which nudges are best suited to helping make people better off as they see the matter. The first of these involves narrowing the definition of appropriate nudging tools, while the second solution requires policy-makers to subject nudges to a behaviorally informed cost–benefit analysis.[11]

Narrowing the nudge definition: the rationality standard

In response to the problem that some interventions can be normatively undesirable even while they nominally preserve choice, a first solution ('the rationality standard') limits the privileged position claimed by nudges to a narrower set of behavioral tools that includes only interventions that are rationality-promoting or rationality-enabling. The former aim to help decision-makers make more rational judgments or decisions, most notably by means of debiasing (Jolls & Sunstein, 2006; Tor, 2008). For instance, if some employees save too little because they underestimate the resources they will need upon retirement, policy-makers can try to foster more realistic assessments of these needs by addressing the sources of the bias. The success of such interventions should lead the newly debiased employees to increase their current savings rates.

11 The following summary descriptions of the two solutions are based on Tor (2016, 2017) and Tor (2015, 2019), respectively.

Notably, those rationality-promoting nudges that policy-makers who are concerned with subjective welfare may adopt are less likely to diminish individuals' subjective well-being than interventions that merely preserve choice.[12] When the former reduce judgment biases or violations of rational choice, they lead to a better alignment of subjective judgments with objective reality or a greater consistency in choice behavior, either of which effect tends to improve individuals' subjective welfare.[13]

On the other hand, a rationality-enabling nudge may involve providing employees with clear information about the many benefits of increased savings or making it easier for them to increase their salary contribution rate. More generally, nudges that do not seek to make people behave more rationally per se can nevertheless make them better off by offering useful information, simplifying their decision-making process or in other ways making the judgment and decision environment within which they operate more hospitable (cf. Klayman & Brown, 1993). By creating a more rationality-friendly environment, these nudges increase the likelihood that individuals' ultimate decisions will improve their subjective well-being.

The rationality standard for nudging is therefore quite attractive. Yet a perusal of the numerous methods of nudging proposed in the literature (Sunstein, 2016) quickly reveals that many of these choice-preserving tools are neither rationality-promoting nor even rationality-enabling. The most problematic of these methods seek actively to diminish rationality. They draw on familiar behavioral phenomena, particularly but not exclusively from the domain of judgmental heuristics, with the goal of shaping choices by making individuals biased (or more biased, as the case may be).

Biasing nudges, for example, may recruit the availability heuristic or anchoring to bias people's judgment in the direction desired by policy-makers. To illustrate, particularly graphic descriptions of the results of insufficient retirement savings or repeated exposure to stories depicting rare cases of extremely

12 While improved rationality can sometimes diminish subjective welfare – as when people are potentially better off retaining their mildly overoptimistic perceptions (Taylor & Brown, 1988) – policy-makers concerned with making people subjectively better off will not seek to implement such nudges.

13 Rationality-promoting nudges will not always benefit their targets in the short term, despite their beneficial tendencies overall. For instance, when individuals act contrary to their subjective well-being due to more than one behavioral factor (e.g., multiple judgment biases or a combination of a bias and a rational choice violation), remediating one source of bias may still fail to change their behavior. Occasionally, moreover, correcting only one of multiple error sources might even lead decision-makers to behave in ways that diminish their subjective welfare (Tor, 2017).

negative outcomes for retirees with insufficient savings may lead employees to overestimate the risks involved in having insufficient savings upon retirement. The nudge would appear successful if it leads these employees to increase their retirement contribution rates. Nonetheless, this apparent success is due to the nudge-induced bias on the part of the employees, who are saving more now only because they overestimate their needs upon retirement.

The success of such biasing nudging would be troubling. Admittedly, it can make subjectively better off those employees who saved insufficiently because of a prior failure of rationality (say, because they were overoptimistic about their expected income stream until retirement), bringing their actual contribution rate closer to what they would have chosen if they were unbiased.[14] At the same time, however, other previously unbiased employees now may increase their contributions excessively, with a concomitant reduction in their subjective well-being.

Yet the harms generated by rationality-diminishing interventions go beyond their complex and sometimes contradictory impacts on the specific behaviors they target. For one, a reduction in the rationality of specific judgment or decision processes may negatively impact individuals' welfare in related areas.[15] An employee who overestimates the risks she is likely to face upon retirement, for example, may go beyond increasing her saving rates and excessively invest in other means of risk reduction, such as paying for objectively unattractive insurance products. In addition to negative spillover effects of this sort, moreover, attempts to make individuals better off by misleading them provide the opportunity for using similar manipulative methods to advance other policy ends beyond means-paternalism. They also risk legitimating such manipulation among policy-makers, opening the door to familiar public choice critiques of behaviorally informed policies. Finally, individuals are likely to be less supportive of policies that seek to bias them compared to policies that use other methods of nudging.[16]

A great many nudges, however, fall somewhere between rationality-diminishing interventions and their rationality-enabling or rationality-promoting counterparts. These common nudges exploit individuals' bounded rationality

14 There is also no guarantee that the contributions of these doubly biased employees will better approximate their behavior if they were unbiased than their pre-nudge biased contribution rates.

15 This concern is distinct from the question of whether diminishing the rationality of one process may negatively affect the target individuals' rationality more generally or create moral hazard problems (e.g., Klick & Mitchell, 2006).

16 These problems may be compounded by significant non-welfarist concerns that are outside the scope of the present analysis.

to accomplish their goals, without diminishing it further. They frequently draw on the various ways in which contextual and situational factors impact choice, including order effects, framing, loss aversion, status quo bias, default effects and more.

Insofar as they do not actively bias their targets' judgment and decision processes, nudges that only exploit individuals' bounded rationality tend to be less harmful than rationality-diminishing ones. But bounded rationality-exploiting nudges still risk some of the harms generated by rationality-diminishing interventions. Most notably, as discussed above, such nudges can be successful even when they are normatively undesirable. Bounded rationality-exploiting nudges can transform the preferences or change the choice behavior of some individuals even when they run contrary to their antecedent preferences, all while providing policy-makers with no evidence that such deleterious effects are occurring.

In addition, efficacious nudges of this sort may also generate some negative spillovers resembling those of rationality-diminishing policies. This is particularly true for interventions that successfully transform preferences (Barnes Truelove *et al.*, 2014; Dolan & Galizzi, 2015). Individuals who follow their new, post-nudge preferences not only reduce their subjective welfare directly (as measured by their antecedent preferences), but also risk distorting their choices in other related contexts. Imagine, for instance, an employee who has been successfully nudged toward weighing their retirement income more heavily compared to present consumption. This employee may well manifest her transformed preferences in other behaviors, perhaps reducing her current consumption beyond what is necessitated by the increased retirement contributions, reducing the risk level of her investment portfolio and so on. However, these post-nudge behaviors all run contrary to her antecedent preferences and therefore lower her subjective well-being.[17]

Policy-makers may also find nudges that exploit people's bounded rationality to be irresistible tools for use in advancing goals beyond subjective welfare, albeit with more limited attendant risks than in the case of rationality-diminishing nudges. For one, those who routinely employ nudges that incidentally transform or override some individuals' preferences may become inured to this concern and thus be more willing intentionally to draw on the same tools to advance paternalistic or social welfare goals that run contrary to their targets' subjective preferences. And, in a similar vein, the employment

17 Spillovers are also possible for interventions that achieve behavioral change while overriding rather than transforming preferences, albeit likely to a lesser extent given the psychological mechanisms that likely produce these effects. For further discussion, see Tor (2019).

of bounded rationality-exploiting nudges also increases public choice concerns more generally.

Whatever its shortcomings, however, nudging that exploits the bounded rationality of its targets also differs from rationality-diminishing interventions in being a seemingly inevitable aspect of the decision environment (or 'choice architecture') in certain instances. *Nudge*'s well-known example of arranging cafeteria shelves in a way that increases the likelihood that children eating there will choose more healthful food options (Thaler & Sunstein, 2008) is a case in point. After all, the hypothetical cafeteria manager must somehow arrange the food shelves and, by assumption, her chosen arrangement will shape the children's choices. Once she is aware of the effect of the arrangement, therefore, the question the manager faces is only how to nudge – that is, how to arrange the shelves – rather than whether to nudge.

Yet the challenge faced by such policy-makers, who cannot avoid design decisions that impact behavior by exploiting individuals' bounded rationality, is not unique. Indeed, whether avoidable or not, a conscientious welfarist would subject all behavioral nudges to some form of a cost–benefit analysis.

A cost–benefit analysis of nudging

In lieu of narrowing the definition of true nudges to exclude choice-preserving interventions that diminish rationality and perhaps even those that merely exploit bounded rationality, the second solution involves subjecting all behavioral policies to a cost–benefit analysis (CBAB). The proposed analysis need not be extensive in every case. Low-risk policies such as rationality-promoting nudges and most rationality-enabling interventions will usually pass the test with only a brief examination, although the substantial difficulties involved in making them effective or finding a practical way to implement them may occasionally still generate costs that exceed their benefits.

On the other hand, the costs of those more problematic forms of nudging that exploit individuals' bounded rationality or even actively diminish rationality should be routinely weighed against their benefits before they are implemented. Notably, the idea of subjecting these nudges to a CBAB runs contrary to one of their important, oft-claimed benefits compared to traditional, costly policy tools – namely, that nudges can generally achieve significant behavior change at a low cost (Thaler & Sunstein, 2008). A related, largely implicit argument is that the costs of nudges are so low compared to their potential benefits that policy-makers need not subject them to the same cost–benefit analysis that is required for determining whether other regulatory interventions are likely to promote social welfare (Adler & Posner, 2006;

Sunstein 2018b). Yet, the preceding discussion made clear that, whatever their benefits, choice-preserving nudges can still impose significant costs on some of their targets.[18]

Proponents of nudging correctly recognize that such policies typically carry much lower direct price tags than those of the more traditional policy tools of economic incentives, mandates or bans (Thaler & Sunstein, 2008; Sibony & Alemanno, 2016). This low-cost intuition stems in part from a focus on the government side of the ledger, which suggests that choice-preserving policies – which by design demand that individuals be free to go their own way if they wish to do so once a nudge is implemented – entail no enforcement costs.[19] Moreover, the targets' freedom of choice also appears to minimize the costs of introducing erroneous nudges, if not for the problems discussed earlier.

In addition, recall that certain nudges involve the unavoidable selection of some form of choice architecture. This is the case, for example, when policymakers must offer some default arrangement, such as an opt-in versus an opt-out option, or even a forced choice among the available options. But the same is true for many other settings that require disclosure or other types of information to flow to individuals, or situations that necessitate the use of forms or procedures as prerequisites for the provision of government services or during other interactions between citizens and public or private institutions. In these and similar circumstances, laws, regulations or institutional arrangements usually exist already, so while potential nudges may require their modification, the direct costs of such changes to extant arrangements tend to be small compared to the overall volume of economic activity they affect.

Nudge advocates do recognize that the development and implementation of efficacious nudges still entail costs (Halpern, 2015). This is, in fact, one important implication of 'Nudges that fail', which concludes that learning to nudge better – namely, the process of experimenting with variants of an attempted nudge until it works as desired – is often an appropriate response to an initial nudge failure (Sunstein, 2017). Still, the direct costs of multiple field studies or even attempted full-scale nudges usually pale in comparison to the potential benefits of a noticeable behavioral change for a substantial population.

18 This summary account does not address the implications of any further benefits or costs nudges may impose on third parties (Tor, 2019).

19 In reality, choice-preserving nudges may entail non-negligible enforcement costs. This may be the case, for instance, when implementation requires the involvement of intermediaries whose interests do not align perfectly with those of the individuals targeted by the policies, as when private companies are required to nudge their employees to increase retirement contributions.

Notwithstanding their relatively limited direct costs for policy-makers, however, some nudges can generate substantial costs for their targets. These costs include, for example, the unintended distortion of some individuals' behaviors following the nudge. Naturally, the risk of distorted behavior is small where rationality-promoting policies are concerned. As noted earlier, most nudges that help people overcome their judgment biases, for instance, increase the likelihood that the resulting choices of these debiased decision-makers will make them subjectively better off. For similar reasons, rationality-enabling nudges also usually do not generate significant behavioral distortion costs.

The same cannot be said, however, of bounded rationality-exploiting nudges. As noted earlier, though these policies do not actively make their targets less rational, they sometimes override or transform antecedent preferences, either of which outcome makes the subject individuals worse off as they see the matter (or as they saw it before the nudge). For example, employees who succumb to a default retirement contribution nudge when they should have opted out will save too much and consume too little of their current income. When such nudge-driven distortions affect only a small proportion of the nudged employees, their ultimate costs will likely be outweighed by the benefits to those other employees who are successfully and beneficially nudged. But when a nudge distorts the choices of a significant fraction of its targets, the benefits it generates for some may be outweighed by the cost to others.[20] Therefore, nudges that exploit bounded rationality will occasionally generate costly distorted choices that policy-makers should take into account.

However, where rationality-diminishing nudges are concerned, the risk of substantial and costly behavioral distortions is even greater. This last and most problematic class of choice-preserving nudges actively introduces biases into their targets' decision processes. Somewhat ironically, to the extent that they fail or impact only a limited fraction of the targeted individuals, these nudges are less problematic than they might have been otherwise. But when and insofar as they are efficacious, such policies risk causing behavioral distortions that exceed those brought about by interventions that only exploit bounded rationality.

In principle, rationality-diminishing nudges may generate immediate benefits that outweigh their costs. An employee who previously saved insufficiently for retirement due to one bias – say, an underestimation of the likelihood she

20 The behavior of individuals who are unaffected by the nudge, on the other hand, is unlikely to generate significant additional costs. Those who opt out when they should bear only the minor opt-out costs, while those who mistakenly opt out when they should have followed the nudge usually are no worse off than they would have been absent the nudge.

would be unemployed for some time prior to her planned retirement – may increase her saving rate toward the appropriate level thanks to a nudge that successfully biases her in a contrary direction (e.g., by causing her to overestimate her retirement needs). In this case, the contrary nudge that intentionally creates a further discrepancy between the employee's judgment and objective reality nevertheless helps better align her contribution rate with what she would have chosen if she were wholly unbiased.

Nonetheless, such beneficial outcomes are exceedingly difficult to accomplish in practice. For one, policy-makers who try to fight bias with bias need to calibrate their newly introduced distortion of their targets' judgments carefully. In the present case, means-paternalists would need both to assess the magnitude of the employee's initial bias and to determine the nature of the intervention necessary to create a countervailing bias of comparable magnitude. Though theoretically possible, these tasks would pose significant challenges even in controlled experimental settings. Where field interventions and the current state of the art are concerned, the likelihood of such calibration is remote. But without calibration even efficacious rationality-diminishing nudges will underperform whenever their effects substantially exceed or fall short of the magnitude of the pre-existing bias whose behavioral effects they seek to negate. Occasionally, such nudges may even be so effective that they lead to a greater, if contrary, distortion of their targets' behavior than the bias they aimed to counter.

Importantly, moreover, even rationality-diminishing policies that are reasonably calibrated risk generating costly behavioral spillovers. As discussed earlier, spillovers can occur when the successful introduction of a rationality-diminishing nudge distorts individuals' behavior in a related domain, leading them to make decisions that are contrary to their subjective preferences. To illustrate, the successfully nudged employee who was led to overestimate her retirement needs may not only save more for retirement (a beneficial outcome, if reasonably calibrated), but also engage in additional, costly acts (e.g., purchasing excess insurance) that make her worse off.

In this respect, therefore, rationality-diminishing interventions tend to be more costly even than their bounded rationality-exploiting counterparts. We saw that the latter tend to diminish the subjective welfare of only that fraction of their targets whose preferences they transform and, sometimes, of those whose preferences these nudges override. However, rationality-diminishing policies bias the judgments of all of the targets they successfully affect. Consequently, the more efficacious these interventions are, the more costly their potential spillover effects.

More generally, this brief discussion suffices to highlight the potential contribution of CBAB as a means for avoiding the adoption of undesirable nudges or

for helping discipline the adoption of behaviorally informed interventions to advance social welfare or traditional paternalistic goals. Policy-makers who find the first, simpler solution – of narrowing the nudge definition to include only rationality-promoting and rationality-enabling nudges – too constraining may thus have an alternative. The path of BCBA is more informationally demanding, requiring assessments of a variety of costs and benefits beyond those the cost–benefit literature currently recognizes (Weimer, 2017). Nevertheless, this second solution may permit even committed means-paternalists the use of at least some bounded rationality-exploiting nudges. And while the employment of rationality-diminishing interventions is less likely to pass muster even on pure welfarist grounds, it remains theoretically possible that certain nudges belonging to this category could survive a careful BCBA, albeit on rare occasions.

Acknowledgments

The author wishes to thank Cass Sunstein, participants at the Notre Dame Law and Market Behavior (ND LAMB) Research Seminar, the 2018 AFED (French Law and Economics Association) Annual Conference and the 2018 Annual Conference of the Spanish Association of Law and Economics, as well as two anonymous reviewers for their insightful comments. Jack Dahm provided helpful research assistance.

References

Adler, M. D. and E. A. Posner (2006), *New Foundations of Cost-Benefit Analysis*, Cambridge, MA: Harvard University Press.

Bar-Gill, O., D. Schkade and C. R. Sunstein (2018), 'Drawing False Inferences from Mandated Disclosures' Behavioural Public Policy, available online at https://doi.org/10.1017/bpp.2017.12.

Barnes Truelove, H., A. R. Carrico, E. U. Weber, K. T. Raimi and M. P. Vandenbergh (2014), 'Positive and Negative Spillover of Pro-Environmental Behavior: An Integrative Review and Theoretical Framework', *Global Environmental Change*, **29**: 127–138.

Beshears, J., J. Choi, D. Laibson and B. Madrain (2010), 'The Limitations of Defaults', Unpublished manuscript, Retrieved from http://www.nber.org/programs/ag/rrc/NB10-02,%20Beshears,%20Choi,%20Laibson,%20Madrian.pdf

Bubb, R. and R. Pildes (2014), 'How Behavioral Economics Trims Its Sails and Why', *Harvard Law Review*, **127**(6): 1593–1678.

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**(3): 1211–54.

Conly, S. (2013), *Against Autonomy: Justifying Coercive Paternalism*, Cambridge, UK: Cambridge University Press.

de Wijk, R., N. Holthuysen, A. Maaskant, I. Polet, E. van Kleef and M. Vingerhoeds, (2016), 'An In-Store Experiment on the Effect of Accessibility on Sales of Wholegrain and White Bread in Supermarkets', *PLOS ONE*, **11**(3), Article e0151915. Retrieved from http://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0151915.

Dolan, P. and M. M Galizzi (2015), 'Like Ripples on a Pond: Behavioral Spillovers and Their Implications for Research and Policy', *Journal of Economic Psychology*, **47**: 1–16.
Glaeser, E. (2006), 'Paternalism and Psychology', *University of Chicago Law Review*, **73**(1): 133–56.
Goldstein, W. and R. Hogarth (1997), 'Judgment and Decision Research: Some Historical Context', in W. Goldstein & R. Hogarth (eds.), *Research on Judgment and Decision Making: Currents, Connections, and Controversies*, Cambridge, UK: Cambridge University Press, 3–65.
Halpern, D. (2015), *Inside the Nudge Unit*, London, UK: WH Allen.
Hausman, D. M. (2012), *Preference, Value, Choice, and Welfare*, New York, NY: Cambridge University Press.
Johnson, E. J. and D. G. Goldstein (2013), 'Decisions By Default', in E. Shafir (ed.), *The Behavioral Foundations of Public Policy*, Princeton, NJ: Princeton University Press, 417–27.
Jolls, C. and C. R. Sunstein (2006), 'Debiasing through Law', *Journal of Legal Studies*, **35**(1): 199–242.
Klayman, J. and K. Brown (1993), 'Debias the Environment Instead Of the Judge: An Alternative Approach to Reducing Error in Diagnostic (and Other) Judgment.' *Cognition*, **49**(1-2): 97–122.
Klick, J. and G. Mitchell (2006), 'Government Regulation of Irrationality: Moral and Cognitive Hazards', *Minnesota Law Review*, **90**: 1620–63.
Lehwinson-Zamir, D. (2015), 'The Importance of Being Earnest: Two Notions of Internalization', *University of Toronto Law Journal*, **65**: 37–84.
Rizzo, M. and D. Whitman (2009), 'The Knowledge Problem of New Paternalism', *Brigham Young University Law Review*, **2009**: 905–68.
Sibony, A. L. and A. Alemanno (2016), 'The Emergence of Behavioural Policy-Making', in A. Alemanno & A. L. Sibony (eds.), *Nudge and the Law: A European Perspective*, Oxford, UK: Hart Publishing, 1–25.
Shafir, E. (ed.) (2013), *The Behavioral Foundations of Public Policy*, Princeton, NJ: Princeton University Press.
Stanovich, K., M. Topiak and R. West (2008), 'The Development of Rational Thought: A Taxonomy of Heuristics and Biases', *Advances in Child Development and Behavior*, **36**: 251–85.
Sunstein, C. R. (2013), 'The Storrs Lectures: Behavioral Economics and Paternalism', *Yale Law Journal*, **122**: 1826–99.
Sunstein, C. R. (2014), 'Nudges v. Shoves', *Harvard Law Forum*, **127**: 210–17.
Sunstein, C. R. (2016), 'The Council of Psychological Advisors', *Annual Review of Psychology*, **67**: 713–37.
Sunstein, C. R. (2017), 'Nudges That Fail', *Behavioral Public Policy*, **1**(1): 4–25.
Sunstein, C. R. (2018a), '"Better Off, As Judged By Themselves": A Comment on Evaluating Nudges', *International Review of Economics*, **65**(1): 1–8.
Sunstein, C. R. (2018b), *The Cost-Benefit Revolution*, Cambridge, MA: MIT Press.
Sunstein, C. R. and R. H. Thaler (2003), 'Libertarian Paternalism Is Not an Oxymoron', *University of Chicago Law Review*, **70**(4): 1159–1202.
Talor, S. E. and J. D. Brown (1988), 'Illusion and Well-Being: A Social Psychological Perspective on Mental Health', *Psychological Bulletin*, **103**(2): 193–210.
Thaler, R. H. and C. R. Sunstein (2008), *Nudge: Improving Decisions about Health, Wealth, and Happiness*, New York, NY: Penguin Books.
Tor, A. (2008), 'The Methodology of the Behavioral Analysis of Law', *Haifa Law Review*, **4**: 237–327.
Tor, A. (2014), 'Understanding Behavioral Antitrust', *Texas Law Review*, **92**(3): 573–667.
Tor, A. (2015), 'The Next Generation of Behavioural Law and Economics', K. Mathis (ed.), *European Perspectives on Behavioural Law and Economics*, Switzerland: Springer, 17–30.

Tor, A. (2016), 'The Critical and Problematic Role of Bounded Rationality in Nudging', in K. Mathis & A. Tor (eds.), *Nudging – Possibilities, Limitations and Applications in European Law and Economics*, Switzerland: Springer, 3–10.
Tor, A. (2017), 'All Nudges Are Not the Same: Why Rationality Matters for Welfare', Unpublished manuscript.
Tor, A. (2019), 'Cost–Benefit Analysis of Behavioral Policies', Unpublished manuscript.
Weimer, D.L. (2017), *Behavioral Economics for Cost-Benefit Analysis: Benefit Validity When Sovereign Consumers Seem to Make Mistakes*, New York, NY: Cambridge University Press.
Willis, L. (2013), 'When Nudges Fail: Slippery Defaults', *University of Chicago Law Review*, **80**(3): 1155–1229.
Zamir, E. (1998), 'The Efficiency of Paternalism', *Virginia Law Review*, **84**: 229–86.
Zamir, E. and B. Medina (2010), *Law, Economics, and Morality*, New York, NY: Oxford University Press.
Zamir, E. and D. Teichman (2018), *Behavioral Law and Economics*, New York, NY: Oxford University Press.