BOOSTS VS. NUDGES FROM A WELFARIST PERSPECTIVE

Till Grüne-Yanoff

Dalloz | « Revue d'économie politique »

2018/2 Vol. 128 | pages 209 à 224

ISSN 0373-2630

Boosts vs. Nudges from a Welfarist Perspective

Till Grüne-Yanoff[a]

This paper compares two kinds of behavioral policies, boost and nudges, with respect to the normative questions they need to answer. Both policies are committed to welfarism – *i.e.* to respecting individuals' subjective reflected attitudes as the basis of judgment about what is good for them. However, because the two policy types affect behavior change in different ways, different normative requirements arise from this commitment. Nudges affect the choice context so as to change behavior, making use of behavioral evidence for stable relations between contextual features and behavioral outcomes. This intervention works irrespective of the nudged individual's understanding, evaluation or participation. Consequently, it is the nudge proponent who must argue that in the planned intervention, the nudge corrects a mistake and leads to a better outcome that is not compromised by the nudging procedure. Boosts, in contrast, affect behavior by training people in the use of decision tools. This intervention works only with the boosted individual's understanding, approval and active participation. Consequently, the boost proponent does not need to answer the difficult normative questions of mistake, welfare improvement or procedural compromise. Although it might be that nudge proponents can answer these questions for many situations, they constitute a normative burden for nudges that boosts can avoid. In this regard, boosts are therefore preferable to nudges.

behavioral policy – nudge – boost – welfare – normativity

Boosts versus Nudges dans une perspective welfariste

Cet article procède à une comparaison de deux formes de politiques comportementales, les *boosts* et les *nudges*, du point de vue des questions normatives auxquelles elles doivent répondre. Ces politiques s'inscrivent toutes les deux dans une perspective welfariste, *i.e.* elles veillent à respecter les attitudes subjectives et raisonnables des individus dans la détermination des jugements concernant le bien-être de chacun d'entre eux. Cependant, dans la mesure où ces deux formes de politiques comportementales affectent les comportements par des biais différents, leur adhésion au welfarisme n'induit pas les mêmes implications normatives. Les *nudges* affectent les comportements en agissant sur le contexte de choix en s'appuyant sur les résultats expérimentaux établissant une stabilité des relations entre les éléments du contexte et les comportements. Ce type d'intervention ne dépend pas du degré de compréhension ou de la participation active et volontaire des individus visés. Par conséquent, la justification des *nudges* repose sur la démonstration que l'intervention va permettre de corriger une erreur comportementale et effectivement mener à un meilleur résultat. Les *boosts*, quant à eux, affectent les comportements en améliorant les compétences des individus à la prise de décision via l'utilisation d'outils. Ce type d'intervention requiert

[a] Royal Institute of Technology (KTH) in Stockholm. Email: gryne@kth.se

la compréhension et la participation active et volontaire des sujets. Par conséquent, la justification des *boosts* ne dépend pas des questions normatives relatives à l'identification des erreurs ou à la détermination de l'amélioration effective du bien-être via l'intervention. Bien que les partisans des *nudges* puissent parfois répondre à ces difficultés normatives, il apparait que les *boosts* les évitent d'emblée. De ce point de vue, les *boosts* apparaissent préférables aux *nudges*.

politique comportementale – nudge – boost – bien-être – normativité

JEL codes: D63, D91

1. Introduction

Nudges are a kind of behavioral policy distinct both from merely informing interventions as well as from incentivizing or coercing interventions, which have become quite popular in policy circles recently (Federal Register [2015], Halpern [2016], OECD [2017]). We have recently argued that nudges are not the only kind of policy occupying this "third ground", and distinguished them from *boosts* as a separate policy kind that affects behavior through different causal pathways (Grüne-Yanoff and Hertwig [2016], Hertwig and Grüne-Yanoff [2017]). In this paper, I will distinguish these two policy kinds further on normative grounds, arguing that their respective commitments to *welfarism* leads to differentiating normative requirements to each.

The basic intuition of welfarism is that for something to be good for a person, it is the reflected attitudes and judgments of this person that constitute this normative judgment. Welfarism is a widely shared view amongst philosophers and economists, and it has recently added new clout to paternalistic policy interventions. In particular, so-called libertarian paternalists justify paternalistic policy interventions (*"nudges"*) because they "make choosers better off, as judged by themselves" (Thaler & Sunstein [2008], 4, *cf.* also 10,12,80).

The assumption of cognitive deficits that is central to libertarian paternalism complicates the realization of welfarist ideals: nudges are supposed to correct cognitive deficits, but for that require information about subjective attitudes unmarred by such deficits. Although I believe that nudges often meet this challenge, it requires them answering difficult normative questions about what constituted the deficit in this particular situation, why the nudged behavior is a welfare-improvement, and why the intervention itself is not welfare degrading.

In the present paper, I argue that *boosts* are more compatible with welfarist views than nudges. While nudges *coopt* people's existing cognitive biases to affect behavioral changes, boosts *train* people in employing existing decision heuristics or employing new ones. Because boosts train competences that agents then choose to apply when they see fit, I argue, they do not require answers to the above difficult normative questions in the same way as nudges do. Thus, for types of situations in which these questions are salient, there is a *prima facie* reason to prefer a boost to a nudge for normative reasons.

The paper is structured as follows. Section 2 sketches the main differences between nudges and boosts. Section 3 describes the welfarist position and argues that nudges are committed to it. Section 4 discusses the three difficult normative questions that nudges must face, and in particular argues that nudge defenders must answer these normative questions for the specific situation in which the nudge is supposed to be implemented. Section 5 argues that boosts do not need to answer these normative questions to the same extent as nudges are required to. Section 6 concludes that boosts are therefore preferable to nudges with respect to these normative commitments.

2. Distinguishing Nudges and Boosts

Boosts and nudges are types of interventions in human deliberation that aim at changing people's behavior in predictable ways. Both nudges and boosts agree that human decision-making is often defective, and that these defects are caused by the employed deliberation processes. That distinguishes them from merely informing interventions, which consider the defect to lie with the inputs to otherwise adequate deliberation processes. Both nudges and boosts also agree that their respective interventions should change behavior without prohibiting options or significantly altering economic incentives, and where this effect is easy and cheap to avoid (Thaler & Sunstein [2008, 6]). That distinguishes them from legal mandates or incentivizing policies.

Boosts and nudges differ, however, in *how* they aim to improve decision-making. While nudges coopt people's existing cognitive biases to affect behavioral changes, boosts train people in employing existing decision heuristics or employing new ones.

The innovative core of the nudge approach is the idea that individuals' cognitive and motivational deficiencies can be harnessed to people's benefit. In the first place, nudge identifies these deficiencies as mistakes or "biases" – as misapplications of cognitive heuristics that might have their justification for some purposes but that yield wrong results when applied to other areas. Examples of such biasing heuristics include taking one's memory of events of a certain class as a representative sample of that class; copying the majority's behavior; or focusing on goals that are realized soon, to the detriment of goals that are realized much later. Nudges aim to correct these mistakes, but not by aiming to eliminate these heuristics or their misapplication directly. Rather, they change properties of the choice architecture – the context in which agents choose – so that the applied heuristics produce more desirable behavior. For example, as it is known that people tend to stick with the default in a choice menu, the nudge approach recommends setting those options as defaults that are considered beneficial for paternalistic or social reasons, for example in retirement plan contribution rate (Beshears *et al.* [2009]), or in organ donation choice (Johnson and Goldstein [2003]). In this sense we say that nudges coopt biasing heuristics for their interventions: they motivate their interventions by identifying mistakes, and then design these interventions on the choice context so that the applied heuristics yield a more desirable result. For the argument in this paper, it is important that both the identification of mistakes and the evaluation of alternative behavior require the nudgers to pass normative judgments.

Boosts, in contrast, aim their interventions at the cognitive heuristics. They aim to improve individuals' skills or decision tools with the purpose of extending the agent's decision-making competences. An example of a boost is training people in using *Simple Rules of Thumb* for financial decision making, without aiming to provide comprehensive accounting knowledge – *e.g.* using a separate drawer for business and household proceeds and writing IOUs for transfers between drawers (Drexler *et al.* [2014]). Another example is training people in *Temptation Bundling*, which helps to overcome self-control problems by coupling instantly gratifying "want" activities (*e.g.*, watching the next episode of a habit-forming television show, checking Facebook, receiving a pedicure, eating an indulgent meal) with engagement in a "should" behavior that provides long-term benefits but requires the exertion of willpower (*e.g.*, exercising at the gym, completing a paper review, spending time with a difficult relative) (Milkman *et al.* [2013]). Boosts thus train people in employing decision heuristics that are better for the given purposes than what people currently use. For the argument in this paper, it is important that the identification of better heuristics requires the boosters to pass normative judgment.

Despite these differences in the causal pathways through which they seek to change behavior, boosts and nudges are often genuine alternative interventions for the same policy goal. If the policymaker for example seeks to increase gym visits, she might either apply a ″reset the default″ (a nudge) or a ″teach a strategy″ boost like temptation bundling.

Nudges and boosts thus are distinct in how they aim to change behavior. Besides differing in various positive assumptions (*cf.* Grüne-Yanoff & Hertwig [2016]) they also differ in the kind of normative judgments they call upon the intervention designer to make. These normative judgments are constrained by their respective commitments to welfarism, as I will argue in the next section.

3. Welfarist Commitments of Nudges

Welfarism is a widely shared view amongst philosophers, economists and political scientists, and it has recently added new clout to paternalistic policy interventions. The basic intuition of welfarism is that for something to be good for a person, it is the reflected subjective attitudes of this person –

what she cares about, what moves her, what she is motivated to seek – that constitute this goodness judgment (Sen [1979]). That these subjective states are supposed to be the result of reflection allows for a certain degree of adjustment – so welfarism does not necessarily tie welfare to only the actual desires of a person. But welfarism requires that "an individual's good must not be something alien—it must be "made for" or "suited to" her" (Rosati [1996, 298]).

The characterizations philosophers have provided for this reflection requirement are rather vague. In particular, it is controversial how much revision is allowed as a consequence of reflection, under rationality and awareness conditions, and when such revision amounts to alienation. Connie Rosati proposes a number of possible formulations of varying strength. The weakest formulation requires that a person must be *capable of caring* about X, for X to be good for a person. The strictest formulation requires that she can care about X *without any marked alteration of her present condition*. In-between these two, Rosati's "two-tier internalism" specifies some conditions under which the person must care about X, and then goes on to require that this person must certify these conditions as relevant for her judgment (Rosati [1996, 307]). For example, I might not care much about my safety when under the influence. However, I consider sobriety a necessary condition for my cares to be normatively relevant; and when sober I very much care for my safety. Therefore, the safety level I care for when sober is to be considered good for me, not the safety level I care for when drunk.

Welfarism in its various forms is a widely shared view amongst philosophers, economists and political scientists. In particular, the view that policy should facilitate optimizing the subjective rather than any purportedly objective welfare measure is the dominant normative position in economics. When assessing comparative goodness of alternative economic arrangements, economists will almost exclusively refer to a welfarist framework: one state of affairs is better for a person than another if and only if she prefers the former to the latter (Mas Collel *et al.* [1995, 80]). The social goodness of that state is then determined as an aggregation (and possible trade-off) between these individual welfare judgments. Because the standard economic approach assumes rationality of individual agents – defined as adherence to preference and belief consistency requirements, self-interestedness and full incorporation of available information – this normative position amounts to a very straightforward and perhaps simplistic version of welfarism: the judgment of what is good for a person is constituted by the actual preferences of that person.

Because they endorse such a straightforward form of welfarism, most economists are opposed to paternalistic interventions. An intervention counts as paternalistic if its main purpose is to improve the state of the agent whose choice the intervention interferes with (Dworkin [2014]). If the goodness of an agent's state is determined by the extent to which it satisfies the agent's preferences, and those preferences are identified through the agents observed choices, then there is no room for paternalism: the agent simply chooses what is best for her, and no intervention in her choices could improve her state.

The behavioral theory that nudges are based on opens up the possibility of paternalistic interventions: it shows that in many situations agents' preferences and beliefs are not consistent and beliefs do not incorporate all available information. Furthermore, it establishes systematic relations between contextual factors and such disturbances of rationality. By pointing to these positive findings alone, without any additional normative assumptions, nudgers can argue that where contexts influence cognitive processes in these ways, observed choices are not necessarily reliable guides to people's subjective attitudes and judgments, and therefore do not automatically satisfy the welfarist criteria. For example, when a strategically placed temptation makes me deviate from my previous plan, when an accidentally set default leads me to stick with that option, or when a carefully formulated advertisement makes me buy a product that I don't want, my resulting choices might not reflect so much my subjective welfare, but rather the influences of these contextual factors extraneous to me. An economist wedded to welfarism would have to concede that paternalist interventions might be welfare improving in these cases.

Early arguments for nudges have proposed arguments of this kind (Sunstein and Thaler [2003], Thaler and Sunstein [2008], Camerer *et al.* [2003]). But nudgers needed to strengthen this argument further in order to make it compatible with welfarism. First, they needed to show that paternalistic intervention was actually called for on a broad front. This went beyond showing that paternalist interventions were possibly welfare-improving in some cases. It required establishing that contextual biases create a systematic deterioration of subjective welfare, rather than the mere possibility that in some cases, subjective attitudes were not realized in actual choice. That is, they needed to argue that contextual factors systematically led people to commit welfare-reducing mistakes. To that end, nudge defenders have argued that contextual influences negatively affect welfare because they cause irrational reasoning and deliberation, for example preference and belief inconsistency and failure to process all available information (Sunstein and Thaler [2003], Thaler and Sunstein [2008]). This argument implies a normative assumption about the welfare-relevance of thus defined irrationalities, which will be discussed in the next section.

Second, they needed to show that the policy result is coherent with the idea of welfarism. Once a mistake in the above sense has been identified, a better alternative needs to be found and the intervention set in such a way that the agent is steered towards that alternative. Yet how can the betterness of the alternative be established in a way that respects the agent's reflected attitudes and judgments? Nudgers have clearly committed themselves to such an welfarist position, claiming that their interventions on people's deliberation lead to choices that "are in their best interest or at the very least are better, by their own lights" (Sunstein and Thaler [2003, 1162-3]) and that "make choosers better off, as judged by themselves" (Thaler & Sunstein [2008], 4, *cf.* also 10,12,80).

Third, they need to show that the way the intervention affects people's behavior does not itself have a negative effect on welfare. For example, interventions due to their very nature might have a negative effect on welfare, as when an intervention is so oppressive that it causes the agent considerable anguish. In the next section, I will discuss the normative assumptions underlying this and the previous arguments.

4. Problems for Welfarist Behavioral Policy

My analysis of the nudgers' welfarist commitments revealed three normative questions: (i) why do the observed contextual influences on behavior constitute welfare-reducing *mistakes*? (ii) How can *better* alternatives to observed choices be identified with reference to welfarist criteria? (iii) What ensures that behavioral interventions themselves do not have a *negative effect on welfare*? Each of these questions poses a challenge to nudge proponents. I discuss their answers and the unsolved problems with these answers in turn.

Question (i) has been addressed by nudge defenders early on. They have argued that contextual factors systematically cause irrational reasoning and deliberation, for example preference and belief inconsistency and failure to process all available information (Sunstein and Thaler [2003], Thaler and Sunstein [2008]).[1] However, it is not trivial to assume that such inconsistencies of subjective attitudes have a relevant negative impact on welfare. In the following I will concentrate on the question whether *preference* inconsistency has a negative welfare effect large enough to justify intervention – but similar arguments can be developed for belief inconsistency.

What is it that makes inconsistent preferences normatively so suspect? Why should one conclude that a cyclical preference between some options, or a overweighing of perceived losses in comparison to gains, or temporally inconsistent preferences cannot express what is good for a person? Such questions might be relatively easy to answer from an purportedly objective welfare perspective: we know that the lives of people with severe self-control issues typically do not go so well, in terms of career, relationships, wealth, etc. But as this paper addresses the issue from an welfarist perspective – a perspective that most economists share, as I argued – one cannot at this point take recourse to non-subjective sources of value. Instead, we need to look at the welfarist arguments against inconsistency in more detail.

The first argument against the normative relevance of inconsistent preferences claims that preference inconsistency purportedly prevents making (consistent) choices (*e.g.* Hausman [2012]). If one cannot base one's choice

[1] Sunstein and Thaler [2003, 1168] for example point out the behavioral evidence "that raised questions about the rationality of many judgments and decisions that individuals make." As examples of such irrationalities, they give violating Bayes' rule, preference reversals, hyperbolic discounting and framing. Based on these findings, they propose "strategies that move people in welfare-promoting directions", arguing "evidence of bounded rationality, and of problems of self-control, is sufficient to suggest that such strategies are worth exploring" (Sunstein and Thaler [2003, 1170]).

on one's preferences (*e.g.* when they are fully cyclical and one chooses by selecting the dominant option), so the argument goes, then such preferences should not form the basis for welfare judgments either. But that is not necessarily true. In conjunction with the appropriate decision process, inconsistent preferences, including cyclical and incomplete ones, can always produce consistent choice. For example, if P is an all-things-considered preference, but is incomplete, a decision process selecting non-dominated alternatives (*i.e.* the ones not dispreferred to any other) is sufficient to produce consistent choice. If P is an all-things-considered preference, but cyclical, a process selecting alternatives that are non-dominated in the subset of all non-cyclical preferences also produces consistent choice (Schwarz [1972]). Consequently, this argument against preference inconsistency fails.

A second argument claims that inconsistent or distorted preferences might produce (inconsistent) choice harmful to people, for example making them exploitable to Dutch Books and Money-Pump schemes (Ramsey [1926], Davidson, Suppes & McKinsey [1955]). Because the inconsistent agents themselves do not desire such harms, their preferences leading to these harms should not be taken as a basis for welfare judgments.

However, the ensuing inconsistent choice might not be harmful for two reasons. First, the decision process, although producing inconsistent choice, might hedge against exploitability. For example, an agent with cyclical preferences who selects the dominant option but never engages in trades of items she previously owned avoids being money-pumped (Cubitt & Sugden [2001]). Second, even if agents do not hedge against such exploitation in their decision mechanisms, they might encounter environments in which such exploitations do not take place – so that no harm ensues from their inconsistent choices. This is a question of *ecological rationality*: it depends on the environment the agent is in and the purposes for which the decision is taken, whether choice consistency matters (Arkes *et al.* [2016]).

Thus, *if* choice consistency is important, then it is important to secure a consistency-producing match between preference and decision process, not consistency in preferences alone. But choice consistency need not even be important, because decision processes might hedge against harm, or the environment is such that incoherence does not matter. So this argument against preference inconsistency also fails.

These arguments show that inconsistency does not necessarily lead to a welfare loss, not that they never do. Thus the justification for a nudge intervention is still possible, but depends on the specifics of the situation: the nudge proponent must show that for this particular situation, the irrationalities caused by the contextual influences indeed have negative welfare effects.

Early nudge defenders largely ignored *question (ii)*. In fact, while expressing their commitment to welfarism, most practical proposals were based on external criteria.[2] More recently, a number of economic authors have pro-

[2] In some cases, the authors link welfare judgments to aggregate data. Claiming that an increase of 401(k) participation would be highly beneficial, for example, they point out that

posed a more thorough treatment under the title *Behavioral Welfare Economics* (Bernheim [2009], Bernheim & Rangel [2009], Manzini & Mariotti [2012] and Rubinstein & Salant [2012]). The main results from this research are a number of strategies that help reconstruct welfare-relevant preferences from context-affected choice data.

Two approaches can be distinguished in this literature: those that maintain the conventional assumption that choice maximizes a single, all-things-considered preference ordering, but argue that these preferences are influenced by additional factors beyond the choice alternatives ; and those that assume some unconventional decision rule connecting preferences and choice (for this distinction, see Bernheim [2009], Rubinstein & Salant [2012], Rubinstein & Salant [2012], Manzini & Mariotti [2014]).

The first approach (also termed the "model-less", "model-free" or the "Pareto approach") distinguishes welfare-irrelevant but choice-influencing conditions ("frames" or "ancillary conditions") that extend choice functions, and infers consistent preference orderings for each such extended choice set. A (potentially incomplete) welfare-relevant "unambiguous" or "union" ordering can be constructed from the intersection of all these preference orderings (examples of this approach are Salant and Rubinstein [2008], Bernheim and Rangel [2009], Apesteguia & Ballester [2010]).

The second approach (also termed the "model-based" approach) posits a cognitive decision process that generates choice from preferences. Preferences then can be identified from inconsistent choice data by abducing them as the initial conditions of the process that produced the choice data. For example, Manzini & Mariotti's [2012] assume that an agent first simplifies her choice set by disregarding some alternatives and then maximizes her preferences over the rest. They take only the latter as welfare-relevant, and seek to identify them from choice data. Most generally, perhaps, Rubinstein & Salant [2012] conjecture that there is a *distortion function* D that attaches to every ordering $>$ the set $D(>)$ of all orderings that may be displayed (through choice) by an individual with the welfare ordering $>$. The task of the welfare theorist is then to extract $>$ from the $D(>)$ expressed in choice (for more examples of this approach see *e.g.* Bleichrodt *et al.* [2001] and Green and Hojman [2007]). In each of these approaches, choice is the product of welfare-relevant preferences distorted through some (welfare-irrelevant) decision process. The reconstruction seeks to extricate the welfare-relevant preference ordering from the observed choice.

But why would these reconstruction strategies yield a *normatively valid* basis for judging certain outcomes as better from a welfarist perspective, than the ones actually chosen? Reconstruction generally does not satisfy the most stringent versions of welfarism, as they require that only the actual attitudes of an agent constitutes the basis for judgments about what is good for her. Instead, reconstructive approaches determine what preferences an

"the U.S. aggregate saving rate is too low" (Camerer *et al.* [2003, 1227]). In other cases, welfare is taken to be the material payoff of an activity, for example the welfare gain of a lottery ticket as "the odds of winning a lottery and of the real payoffs in terms of the after-tax discounted present value of earnings" (Camerer *et al.* [2003, 1231]). For further discussion, see Grüne-Yanoff [2012].

agent would have had under counterfactual conditions. Results of reconstructive approaches might however satisfy weaker versions of welfarism, because these versions allow for preference determination under counterfactual conditions. To give just one example, Rosati's "two-tier internalism" constrains the allowed counterfactual conditions as those that the person in question herself certifies as relevant for her judgment. The question that needs to be addressed, then, is whether the reconstructive approaches satisfy these constraints.

My argument is that while they might under some circumstances satisfy such constraints, it is not the case that they generally do. This is particularly clear with respect to the model-based reconstruction strategy. First, it assumes that the real cognitive process only has a distorting influence and is itself welfare-irrelevant. In contradiction to this assumption it seems possible that the process is welfare-relevant, either by expressing welfare-relevant attitudes (*e.g.* lexicographic preferences expressed through a lexicographic decision rule, Mandler *et al.* [2012]), or by deriving choice from welfare-relevant but non-standard preferences in a welfare-relevant way (as in the decision rules dealing with incomplete or cyclical preferences). Some defenders of the model-based approach acknowledge this possibility themselves:

> "that the categorization process itself contains welfare-relevant aspects (if for example a jam brand X is favored over another brand Y it may be the case that on average brand X jams prove to be better than brand Y jams)" (Manzini & Mariotti [2014, 350])

Thus, whether a model-based reconstruction yields a welfare-relevant basis or not will depend on the particular circumstances under investigation, viz. the preferences to be reconstructed and the assumed decision processes.

The same holds for the model-free approach. In particular, it assumes that the distorting effects of various decision processes show themselves in the context-dependence of the resulting choice. This is why model-free approaches differentiate choices by choice set and ancillary conditions, and then identify those choices as welfare-relevant that turn out to be stable across such contexts. But it isn't obvious that distortions are always context dependent. Consider for example the decision rule to always choose preference-dominated options. It is plausible to consider such a rule as being welfare-reducing, as the agent following it will always choose what is worse for her. But because she will *always* do so, choices based on that rule will be stable across ancillary conditions, and the model-free approach will not remove it from the set of welfare-relevant choices. Thus the model-free approach, like the model-based approach, requires further normative considerations of the particular circumstances in order to judge whether a reconstruction along their lines is a normatively valid basis for welfare judgment. This is perhaps not a problem for the reconstructive approaches, but it shows that no general conclusion about the normative adequacy of these reconstructive strategies can be drawn.

Question (iii) asks whether the implementation of an intervention itself could possibly have a negative effect on welfare. For example, an intervention might be so oppressive that it causes the agent considerable anguish.

Similarly, loss of autonomy or the feeling of being manipulated might reduce the positive effects of otherwise welfare-enhancing interventions.

This question has received some attention from early nudge defenders. Sunstein and Thaler even suggest that reduction in liberty through behavioral intervention should be integrated into the overall welfare assessment of potential policies (Sunstein and Thaler [2003], footnote 22). Such a perspective might be feasible, but only after the welfare-detrimental effects of an intervention have been determined.

Potential welfare-detrimental effects that both defenders and critics of nudges have discusses include manipulation and non-transparency (Sunstein [2016], Hausman and Welch [2012]). If an agent feels manipulated into choosing an otherwise good option, or feels that the intervention was not transparent to her, this feeling might substantially reduce the welfare derived from having chosen that option.

An intervention is manipulative if it bypasses or subverts the rational capacities of the person being influenced (Wilkinson [2013]). An intervention is non-transparent if it is implemented in a way that makes it difficult for the affected agent to learn about this intervention and its effects. Whether a behavioral intervention is manipulative or non-transparent thus depends on the mechanisms through which the intervention operates. Yet for many nudge interventions, multiple mechanistic explanations have been proposed. Take the example of default-setting, a popular nudge intervention: it might sometimes operate through cognitive effort avoidance, sometimes through a recommendation effect, and sometimes through loss aversion (Grüne-Yanoff [2016]). These mechanisms might operate side-by-side in different members of the same population, but it is plausible to assume that the frequency with which they are found to operate in a population is influenced by contextual factors. Consequently, to determine whether (or to what extent) a behavioral intervention was manipulative or non-transparent, it has to be determined through which mechanism that intervention operated – which in turn depends on contextual factors. To answer the question whether the implementation of an intervention itself has a negative effect on welfare thus requires an investigation of the specific context of implementation.

5. Why These Problems do not Arise for Boosts

In the previous section, I argued that in order to answer three crucial normative questions about nudges, the specific context in which the intervention is supposed to be implemented must be investigated. These three questions were: (i) whether a specific behavior constituted a mistake, (ii) what constituted a better alternative behavior, and (iii) whether the proposed intervention might be welfare-detrimental. Each of these questions must be answered for each nudge intervention. If they are not, it is possible that the nudge yields a welfare loss and thus is not justified from a welfarist perspective.

In this section, I show that arguing for the implementation of a boost in a particular situation does not require answering these three questions in the same way as the nudge proponent is required to. Nudges thus have to address and deal with an additional *normative burden* that boosts can avoid. Three features of boosts are responsible for this difference. First, an implementation of an effective boost only provides a strategy for behavioral change – in contrast to the implementation of an effective nudge, which directly affects behavior in a certain way. Second, an effective boost requires the participation of the boosted agent, while a nudge can often be effective without the nudged agent's active participation. Third, effective boosts are necessarily transparent, while effective nudges need not be. I will now explain these differentiating features in more detail and show how they reduce boosts' normative burden.

Boosts aim to change behavior by intervening on agents' cognitive heuristics. In the first place, this requires that boost proponents identify a deficit: for example, that people apply heuristics that yield less than optimal behavioral results. Without such an argument, developing and implementing a boost would be unmotivated. In this, they start out similarly to nudge proponents. But such a motivating identification is *general* and *conjectural*: as a reason for a boost, it suffices to argue (even with low confidence) that some people in some situations might make such a mistake. For the nudge proponent, this is not sufficient: she must argue (with high confidence) that *for the particular population* for which the intervention is proposed, people make this mistake. For if this were not the case, an effective nudge would change people's behavior, although there was nothing wrong with that behavior (of those people, in that context) in the first place. Boosts do not face this problem: an implementation of an effective boost only offers a strategy for behavioral change – in contrast to the implementation of an effective nudge, which necessarily effects behavior in a certain way. Boosts train people in more effective heuristics, but leave it to individual agents when to apply these heuristics. It is thus the individual's responsibility, and not the boost proponent's, to assess whether she uses a suboptimal heuristic in a particular context. Thus boosters, in contrast to nudgers, can avoid the difficult question whether particular people in a particular situation systematically commit mistakes or not.

Boosts, like nudges, do not only aim to steer people away from mistakes, but towards better alternative behavior. In particular, they aim to improve individuals' skills or decision tools with the purpose of extending the agent's decision-making competences. This requires that proponents of a boost show that such a skill or tool indeed constitutes an improvement or competence increase. Nudge proponents might claim that such an argument puts boosts in the same normative predicament as nudges, which need to show that the agents who have been effectively nudged into a different behavior are now better off than before. But that is not true.

Nudgers make use of their knowledge of a stable relation between contextual features and behavioral outcomes, and adjust the context so as to change behavior. For each particular intervention, they thus need to show that the induced behavior is indeed better than the pre-nudge alternative, with the additional difficulty that they need to show this with a subjective welfare criterion. Boost proponents do not need to make that argument. Once they established that a particular training intervention could *in some cases* improve competences, they are justified in training people. The application of the boosted heuristic to a particular situation, however, requires the participation of the boosted agent – in contrast to the nudge, which often is effective without the nudged agent's active participation. It is thus the individual who determines whether the application of this heuristic yields a better result. This both satisfies the welfarism condition and absolves the boost proponent from making the difficult welfare judgment for particular cases. Thus boosters, in contrast to nudgers, also avoid the normative burden in this regard.

Finally, there is the worry that the behavioral intervention itself is detrimental to welfare. Even nudges that correct a clear mistake and that steer the nudged to obviously better behavior might fail in this regard – if the nudge were very highly manipulative, or oppressive, for example, then the negative effects of the policy procedure might overwhelm the positive effects of its behavioral result. Importantly, nudges might have these procedural disadvantages even though the nudged might not be aware of them, at least not at the time of implementation. One can imagine a situation in which an individual is nudged to make much higher contributions to her retirement savings, and is entirely satisfied with the changes in her behavior. Later, however, she finds out that the nudged employed manipulative strategies, which she judges so abhorrent that overall, she considers herself worse off being nudged than not. Thus because nudge procedures might have negative welfare effects on the nudged, and because nudges are not necessarily transparent to the nudged, the nudge proponent must argue for the normative acceptability of the nudge before is implemented.

The boost proponent, in contrast, is not required to make this argument. Boosts require a higher degree of transparency than nudges – otherwise, they could not suppose the participation of the boosted agent and thus would inevitably fail. For example, a nudge that changes a default might effectively change behavior despite the nudgee being entirely unaware of the default, the intervention or its effect on her. In contrast, a boost is effective only if an individual (i) is taught a heuristic, and (ii) is trained in applying this heuristic in certain contexts. This does not require that the boosted individual has a full understanding of how and why the heuristic works, but she must at least be able to identify the dominant applicability conditions. This higher transparency requirement, however, makes it very unlikely that the boosted individual later will consider the boost to be so manipulative or oppressive as to have a negative welfare effect. Consequently, boosters face a lesser argumentative burden than nudgers for the normative acceptability of their interventions.

6. Conclusion

In this paper, I compared two kinds of behavioral policies, boosts and nudges, with respect to their respective normative commitments. The nudge concept is committed to welfarism, and due to the way it is implemented must answer three difficult normative questions: what constituted the mistake in particular situations, why a nudged behavior is a welfare-improvement, and why a nudge intervention itself is not welfare degrading. Boosts do not face these normative questions with equal urgency, because they only provide heuristics for behavioral change, because they require the boosted individual participation in its implementation, and because they necessitate a higher level of transparency than nudges to be effective.

This does not imply that boosts are normatively preferable to nudges in *all* situations – in a specific context, a particular nudge might be normatively preferable to any available boost. However, for types of situations in which any of the three normative issues are salient, there is a *prima facie* reason to prefer a boost to a nudge for normative reasons. My argument thus point to the need of developing a typology of situations, in order to more clearly identify *when* these normative questions are salient, and thus when one should expect to be boosts less normatively problematic than nudges. This paper thus continues a project begun for questions of *policy effectiveness* by Grüne-Yanoff, Marchionni and Feufel [2018] and Hertwig [2017], and is now expanded to questions of normative acceptability.

References

APESTEGUIA J., and BALLESTER M. A. [2015], A measure of rationality and welfare. *Journal of Political Economy* 123(6), 1278-1310.

ARKES H. R., GIGERENZER G. and HERTWIG R. [2016], How bad is incoherence? *Decision* 3, 20-39.

BERNHEIM B. D. [2009], Behavioral welfare economics. *Journal of the European Economic Association* 7(2-3), 267-319.

BERNHEIM B. D. and RANGEL A. [2009], Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics* 124, 51-104.

BESHEARS J., CHOI J. J., LAIBSON D., and MADRIAN B. C. [2010], The impact of employer matching on savings plan participation under automatic enrollment. In D. A. Rise (Ed.), *Research findings in the economics of aging* (p. 311-327). Chicago, IL: University of Chicago Press.

BLEICHRODT H., PINTO-PRADES J. L. and WAKKER P. [2001], Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science* 47, 1498-1514.

CAMERER C., ISSACHAROFF S., LOEWENSTEIN G., O'DONOGHUE T. and RABIN M. [2003], Regulation for Conservatives: Behavioral Economics and the Case for Asymmetric Paternalism. *University of Pennsylvania Law Review* 151(3), 1211-1254.

CUBITT R. and SUGDEN R. [2001], On money pumps. *Games and Economic Behavior* 37, 121-160.

DAVIDSON D., MCKINSEY J. and SUPPES P. [1955], Outlines of a formal theory of value, I. *Philosophy of Science* 22, 140-160.

DREXLER A., FISCHER G. and SCHOAR A. [2014], Keeping it simple: Financial literacy and rules of thumb. *American Economic Journal: Applied Economics* 6, 1-31.

DWORKIN G. [2014], Paternalism. In *The Stanford Encyclopedia of Philosophy (Summer 2014 Edition)*. Edward N. Zalta. http://plato.stanford.edu/archives/sum2014/entries/paternalism/.

GREEN J. R. and HOJMAN D. A. [2007], Choice, Rationality and Welfare Measurement. *Harvard Institute of Economic Research Discussion Working Paper Series*, No. 2144.

Federal Register [2015], Executive Order 13707 "Using Behavioral Science Insights to Better Serve the American People". Signed: September 15, 2015. 80 FR 56365, September 18, 2015

GRÜNE-YANOFF T. [2012], Old Wine In New Casks: Libertarian Paternalism Still Violates Liberal Principles. *Social Choice and Welfare* 38(4), 635-645.

GRÜNE-YANOFF T. [2016], Why behavioural policy needs mechanistic evidence. *Economics & Philosophy* 32(3), 463-483.

GRÜNE-YANOFF T., & HERTWIG R. [2016], Nudge versus boost: How coherent are policy and theory? *Minds and Machines* 26, 149-183.

GRÜNE-YANOFF T., MARCHIONNI C. and FEUFEL M. [2016], *The ecological rationality of behavioural policies: How to choose between boosts and nudges.* Manuscript submitted for publication.

GRÜNE-YANOFF T., MARCHIONNI C. and FEUFEL M. [2018], Toward A Framework For Selecting Behavioural Policies: How To Choose Between Boosts And Nudges. *Economics & Philosophy, in press.*

HALPERN D. [2016], *Inside the Nudge Unit: How small changes can make a big difference*. Random House.

HAUSMAN D. M. [2012], *Preference, value, choice, and welfare*. Cambridge, United Kingdom: Cambridge University Press.

HAUSMAN D. M. and WELCH B. [2010], Debate: To Nudge or Not to Nudge. *Journal of Political Philosophy* 18(1), 123-136.

HERTWIG R. [2017], When to boost? Rules for policymakers. *Behavioural Public Policy*, in press.

HERTWIG R. [2017]. When to consider boosting: Somes rules for policy-makers. *Behavioural Public Policy, 1, 143-161.*

HERTWIG R. and GRÜNE-YANOFF [2017]. Nudging and Boosting: steering or empowering good decisions. *Perspectives on Psychological Science 12: 973-986.*

HERTWIG R. and GRÜNE-YANOFF T. [2017], Nudging and Boosting: Two Distinct Pathways to Behavior Change. *Perspectives on Psychological Science*, in press.

INFANTE G., LECOUTEUX G. and SUGDEN R. [2016], Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioral welfare economics. *Journal of Economic Methodology* 23(1), 1-25.

JOHNSON E. J., & GOLDSTEIN D. [2003], Do defaults save lives? *Science* 302(5649), 1338-1339.

MANDLER M., MANZINI P. and MARIOTTI M. [2012], A million answers to twenty questions: Choosing by checklist. *Journal of Economic Theory* 147(1), 71-92.

MANZINI P. and MARIOTTI M. [2012], Categorize then choose: Boundedly rational choice and welfare. *Journal of the European Economic Association* 10(5), 1141-1165.

MANZINI P. and MARIOTTI M. [2014], Welfare economics and bounded rationality: the case for model-based approaches. *Journal of Economic Methodology* 21(4), 343-360.

MAS-COLELL A., WHINSTON M. D. and GREEN J. R. [1995], *Microeconomic theory* (Vol. 1). New York: Oxford University Press.

MILKMAN K. L., MINSON J. A. and VOLPP K. G. [2013], Holding the hunger games hostage at the gym: An evaluation of temptation bundling." *Management Science* 60 (2), 283-299.

OECD [2017], *Behavioural Insights and Public Policy. Lessons from Around the World.* OECD Publishing.

RAMSEY F. P. [1928], Truth and Probability, in *The Foundations of Mathematics and other Logical Essays*, ed. R. B. Braithwaite, London: Routledge & Kegan Paul, 1950.

ROSATI C. [1996], Internalism and the good for a person. *Ethics: An International Journal of Social, Political and Legal Philosophy* 106(2), 297-326.

RUBINSTEIN A. and SALANT Y. [2012], Eliciting welfare preferences from behavioral data sets. *The Review of Economic Studies* 79(1), 375-387.

SALANT Y. and RUBINSTEIN A. [2008], (A, f): choice with frames. *Review of Economic Studies* 75, 1287-1296.

SCHWARTZ T. [1972], Rationality and the Myth of the Maximum. *Noûs* 6, 97-117.

SEN A. [1979], Utiliarianism and Welfarism. *The Journal of Philosophy* 76, 463-489.

SUNSTEIN C. R. [2016], *The ethics of influence: Government in the age of behavioral science*. Cambridge: United Kingdom: Cambridge University Press.

SUNSTEIN C. R., and THALER R. H. [2003], Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review* 70(4), 1159-1202.

THALER R. and SUNSTEIN C. R. [2008], *Nudge: Improving decisions about health, wealth and happiness*. New York, NY: Simon and Schuster.

WILKINSON T. M. [2013], Nudging and manipulation. *Political Studies* 61(2), 341-355.