Economics and Philosophy, **35** (2019) 195–222 © Cambridge University Press
doi:10.1017/S0266267118000196 First published online 23 August 2018
cambridge.org/eap

NUDGE VERSUS BOOST: A DISTINCTION WITHOUT A NORMATIVE DIFFERENCE

ANDREW SIMS*, THOMAS MICHAEL MÜLLER†

Abstract: Behavioural public policy (BPP) has come under fire by critics who claim that it is illiberal. Some authors recently suggest that there is a type of BPP – boosting – that is not as vulnerable to this normative critique. Our paper challenges this claim: there's no non-circular way to draw the distinction between nudge and boost that would make the normative difference required to infer the permissibility of a policy intervention from its type-membership. We consider two strategies: paradigmatic examples and causal mechanisms. We conclude by sketching some suggestions about the right way to approach the normative issues.

Keywords: behavioural public policy, autonomy, libertarian paternalism, nudge, bounded rationality

1. BEHAVIOURAL PUBLIC POLICY IN LIBERAL DEMOCRACIES

Behavioural public policy (BPP) is an approach to state intervention that uses knowledge about the bounded rationality of human reasoners – largely from cognitive psychology and behavioural economics – in order to achieve policy aims (Thaler and Sunstein 2008). The approach starts out from results in these disciplines that appear to demonstrate that human beings systematically depart from the norms of economic rationality in a range of experimental contexts. These departures have been shown in a number of different cases: they include a tendency in statistical reasoning

* Institut supérieur de philosophie, Place Cardinal Mercier 14, Ottignies-Louvain-la-Neuve 1348, Belgium. Email: andrew.sims@uclouvain.be. URL: https://uclouvain.academia.edu/AndrewSims.

† Laboratoire d'Économie Dionysien (LED), Rue de la Liberté 2, 93526 Saint-Denis, France. Email: thomas.muller@univ-paris8.fr. URL: https://sites.google.com/site/up8led/members/internal-members/hpe/mueller-thomas.

to ignore base rates, a tendency to disproportionately value small and immediate rewards over large rewards that are relatively distant in time, a tendency towards an 'inertia bias' in decision making when given a default option, a tendency to make decisions on the basis of irrelevant information associated with the way that some choice is framed, and so on (Kahneman 2011).

Classical approaches to policymaking in liberal democracies often assume the basic rationality of the self-interested individual, and that the individual will be successful in satisfying her preferences in the absence of impediment. But these recent results suggest that individuals may actually have a hard time satisfying their preferences, or in making their preferences consistent with their values or beliefs about the good. For example, people with poor eating habits sometimes both agree that they should improve their behaviour and form an intention to do so, but then have trouble in successfully implementing this intention. Thaler and Sunstein (2008) have therefore proposed that states could intervene to improve outcomes by using knowledge about bounded rationality to act upon the context in which people make choices. Specifically, by altering the 'choice architecture' in order to recruit the aforementioned heuristics and biases, policymakers can harness the irrationality of individuals on a large scale in order to nudge them towards the behaviours which would be chosen under ideal circumstances: complete information, unlimited cognitive resources, and efficacy of will.

A paradigmatic example is the 'Save More Tomorrow' programme (Thaler and Benartzi 2004). Many employees do not save enough money to acquire a comfortable retirement fund, and this intervention is intended to ameliorate that deficit. The way it works is by presenting the choice to save money in a way that draws upon knowledge of loss aversion and temporal discounting. If the choice to sacrifice money to a retirement fund is presented in terms of an immediate loss, then agents are less likely to do so because they over-value those immediate losses in a way that is distorted against the norms of economic rationality. In this context, Save More Tomorrow intervenes by presenting the choice to save in a way so that the loss will take place at a future moment – money is deducted when the employee receives a wage raise. It thereby corrects for the evaluative distortion, and appears to have been successful.[1]

There is a normative controversy associated with BPP which will be our primary focus in this paper. That is whether BPP is somehow manipulative or paternalistic in a way that is incompatible with the normative commitments of classical liberalism. This is important

[1] Save More Tomorrow also relies on setting itself as the default option, so that participants need to opt-out. For simplicity's sake we have just focused on the component based on loss aversion as a general illustration of what BPP involves.

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

because these commitments are shared by the Western governments that implement those policy interventions. Proponents of BPP often claim that it implies a political philosophy – libertarian paternalism – that is compatible with liberal values. That is because BPP does not coerce the individual into making any particular choice, nor does it significantly change the incentive structure for the individual through financial rewards or penalties. Rather, it just changes the way in which choices are presented to individuals, so that they are more likely to choose the way they would under ideal conditions. For example, in Save More Tomorrow, and although the presumption is made that the individual would prefer to save more for retirement, the individual may nonetheless make the decision *not* to save if she so wishes. In this sense the individual is not deprived of her liberty, and retains the ability to do otherwise. However, this assessment is not shared by all authors. For instance, Grüne-Yanoff (2012), a prolific commentator on these issues, has argued that BPP has the potential to be illiberal if it fails to take into account the plurality of values across individuals. That is because BPP, at least in the form originally endorsed by Sunstein and Thaler (2008), is uniform in its application to individuals. The major proponents of BPP continue to protest mightily against such critiques (e.g. Sunstein 2015).

Now, it's probably the case that this is a question that will not admit of a categorical answer. By this we mean that it's unlikely that BPP is either always illiberal or always liberal: the truth will probably lie somewhere in between. But then we need a framework or strategy to help us distinguish BPP interventions that are liberal from those that are not. This paper has to do with one kind of strategy for addressing this controversy: make type-distinctions between BPPs in a way such that we can give answers to the liberality question for particular policy interventions just on the basis that they belong to one of the distinguished types. Call this the *type-distinction strategy*.

The type-distinction strategy is ambivalently endorsed by Grüne-Yanoff and Hertwig (2016), who contrast the BPP types of *nudging* and *boosting*, partly in order to suggest a resolution to this issue:

> nudgers assume homogenous approximations of welfare-enhancing goals, thus easily turning nudges into standard paternalistic interventions that are inconsistent with the autonomy-improvement argument. … In contrast, boost policy designers seek to improve people's capacity to reach certain goals, but leave the choice of means to the decision makers themselves. (Grüne-Yanoff and Hertwig 2016: 177–178)

The general idea is this: if we can distinguish between types of BPP having properties implying different answers to associated normative questions – for example, if they necessarily differ along the dimensions of transparency and respect for autonomy (Hertwig and Grüne-Yanoff

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

2017: 9–10) – then we can give answers about the liberality of a given intervention based on membership to that type rather than on an individual evaluation of the particular policy intervention. In this context, Grüne-Yanoff and Hertwig (2016) have suggested that we can contrast nudge policy interventions (already long discussed after being introduced in Thaler and Sunstein 2008) and boost policy interventions, and that this contrast will be able to do that work: 'boosts require minimal competences and motivations on the part of the target audience to be effective. Consequently, the criticism that nudge policies infringe on human autonomy and dignity does not apply (or applies less) to boost policies' (Grüne-Yanoff and Hertwig 2016: 176).

Sadly, however, we don't think that the distinction between nudge and boost *can* do this work. Let's be clear about what we are claiming here. First, we are not claiming that the distinction between nudge and boost marks no difference whatsoever. We believe that Grüne-Yanoff and Hertwig are on to something important, and that the distinction has a significant heuristic value in suggesting novel interventions to policymakers. What we are denying is that the distinction marks a normative difference significant enough to underwrite a type-distinction approach to the controversy about which BPP interventions are permissible. Second, we do not think that endorsement of the distinction between nudge and boost in a weaker and more pragmatic sense commits one to thinking that it marks a normative difference. It appears to us that in their discussion of this question Grüne-Yanoff and Hertwig equivocate between a weaker and more pragmatic position on which the distinction is merely suggestive and a stronger and more substantial position on which it marks a genuine normative difference. (This equivocation can be seen in the quotation above – does the normative criticism not apply, or does it just apply 'less'?) So it may be a result of our criticism that these authors wish to retreat to the more pragmatic position, and we would have no issue with that. However, it would mean that they are not entitled to make any normative claims about a BPP intervention on the basis of its type membership – not even that normative criticisms apply to the type 'less'. That would just be a sort of empirical claim in need of evidence: that numerically speaking there are more permissible boost policies than there are permissible nudge policies.

Here is how we argue our claim. First, in Section 2, we motivate the distinction between nudge and boost in some detail and give examples for the sake of illustration. Then we give a more detailed explanation of how the distinction may do work in resolving the normative controversy around BPP. Then, we present the main argument to the effect that the distinction between boost and nudge does not mark a significant normative difference. First, in Section 3, we show that the paradigmatic examples of boost which are used to set them apart as

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

a type are systematically underdetermined, and furthermore that it is unlikely that one can infer the relevant normative properties from the descriptive properties that are suggested as characteristic (like the need for cooperation). We are therefore in agreement with Grüne-Yanoff and Hertwig (2016) when they advocate the need for a mechanistic criterion to mark the distinction. To that end, in Section 4 we evaluate this approach. Without ruling out future success in this area, we conclude that current attempts are inadequate. Last, in Section 5, we suggest the outlines of an alternative methodology for appraising individual BPP interventions with respect to liberality: our suggestion is that this needs to be done on a case-by-case basis. We briefly consider the possibilities that lie within this area.

2. THE DISTINCTION BETWEEN NUDGE AND BOOST

Bounded rationality is a concept introduced as such by Herbert Simon (1956). The insight that motivates the introduction of this concept is that cognitive processes are always constrained by a number of factors that make context-neutral optimization impracticable – factors like limitations of time, capacity, and the difficulty of the problem. So, the line of thought goes, we would do well not to theorize cognition as a process for producing context-neutrally optimal solutions with respect to some set of problems. Instead, we should theorize it as a process for producing sufficient or *good enough* solutions to that set of problems, given a particular environmental context.

Different authors have developed this insight in different directions, which has generated distinct research programmes. In order to introduce the distinction between nudge and boost, Grüne-Yanoff and Hertwig (2016) discuss two of these research programmes: 'Heuristics and Biases' and 'Simple Heuristics'. Heuristics and Biases conceptualizes bounded rationality in terms of a set of systematic biases which lead to a consistent departure from coherence-based norms of rationality. These departures are seen as evidence that human beings are constitutionally flawed reasoners. The Heuristics and Biases approach has been brought to great prominence by the pioneering work of Tversky and Kahneman (1974) and popularized in a wider context by Kahneman (2011). It exerts further influence in dual-process theories of cognitive architecture (Evans and Stanovich 2013).

Now, Hertwig and Grüne-Yanoff (2017) discuss a number of other research programmes for bounded rationality in psychology – such as 'naturalistic decision making'. It's plausible to think that these could themselves give rise to distinct styles of behavioural public policy, as these authors themselves note. However, these authors also choose to limit themselves to Simple Heuristics in the elaboration of their vision for policymaking. For that reason we also limit ourselves to discussing

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

this approach. To cover every single permutation at length would be impossible in a single article. In the absence of a good argument to the contrary, we assume that our arguments for the separation of descriptive and normative concerns apply just as well to other policy approaches.

The Simple Heuristics programme is championed by Gigerenzer et al. (1999). Simple Heuristics emphasizes the importance of an adaptive fit between the reasoner and her environment. Proponents often stress that our reasoning capacities are a form of ecological rationality, in that they rely on a structural isomorphism between the reasoning process (e.g. heuristics) and the environment. Therefore, reasoners may perform more or less sufficiently depending on how the reasoner's environment is structured. Furthermore, they stress that the fast and frugal heuristics that we rely on to solve problems may – under the right conditions of good fit between agent and world – be even more efficient than an approach following coherence-based norms.

These two research programmes inspire distinct approaches to BPP. These are the nudge approach and the boost approach. Nudge interventions follow Heuristics and Biases in the way that they conceive of bounded rationality as imperfect rather than ecological, and seek to recruit these same imperfections in order to bring actual reasoning into line with the coherence-based norms of economic rationality: they seek to make the behaviour of individuals as close to economically rational as possible, under the assumption that economically rational decision-making will result in better outcomes in the relevant domains (Thaler and Sunstein 2008).[2] This need not involve making individuals more competent reasoners; it can steer people's decision-making in a covert manner.

Boost interventions follow Simple Heuristics in their ecological interpretation of bounded rationality. This ecological interpretation allows policymakers to intervene in two distinct ways: (i) by altering the structure of the environment so that it is better suited to the individual's heuristics; or (ii) educating the individual in new heuristics that are better suited to common environments or problems, thereby boosting their 'heuristic repertoire' (Grüne-Yanoff and Hertwig 2016; Hertwig and Grüne-Yanoff 2017). The idea is that this will allow individuals to better satisfy their existing preferences, having 'the goal of empowering people by expanding (boosting) their competences and thus helping them to reach their objectives (without making undue assumptions about what those objectives are)' (Grüne-Yanoff and Hertwig 2016: 156).

[2] There are complications with this broad characterization, of course. For example, so-called prosocial nudges aim to promote public goods, and some so-called nudges do not recruit heuristics and biases at all. This gives rise to terminological issues. But it's clear that the type we describe is considered the most controversial and interesting in the literature. A nice typology of putative nudges is sketched in Barton and Grüne-Yanoff (2015).

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

There is an initial contrastive set of examples that we can give at this point. The examples are not intended to exhaustively characterize either approach, nor to provide the basis of an argument that they *cannot* be so characterized, but just to serve two roles: (i) make the distinction intuitively clearer in a more concrete context; and (ii) show why it is superficially promising as a basis for answering the normative question about BPP. Both examples are given by Grüne-Yanoff and Hertwig (2016), which is the first systematic attempt at making the distinction, though the first of these examples originates in Thaler and Sunstein (2008).

The example of a nudge occurs in the context of choice-framing. Nudge interventions sometimes draw upon framing effects – the influence of irrelevant contextual features of a given choice – in order to have the targets of the intervention choose the 'best' option. For example, it is well known that actual reasoners are loss-averse: they are more likely to eschew a course of action if the associated potential for loss is made salient. Therefore, if a government wants to promote a particular energy conservation method they can frame the choice to adopt it in terms of loss rather than gain:

> (Loss) If you do not use energy conservation methods, you will lose $350 per year.
> (Gain) If you use energy conservation methods, you will save $350 per year.

This is an example of a nudge because it alters the choice architecture in such a way that will steer the individual towards a decision that it is assumed they would prefer (to save money at little to no cost). It does so by exploiting the documented loss-aversion of human reasoners, and not through changing the incentives through rewards or punishments. Furthermore, although it exerts influence in a way that bypasses the rational faculties of the individual, it does not do so in a way that prevents her from doing otherwise.

This example of a nudge can be contrasted with a boosting policy intervention that illustrates both of the ways in which a boost can take place (competence-boosting and environmental restructuring). This intervention was based on the finding that individuals are much better at statistical inference when the relevant statistical information is represented in the form of frequencies rather than in the form of percentages or probabilities (Gigerenzer and Hoffrage 1995). For example, medical doctors are more likely to make better inferences (in the case below, estimating the chance that a woman with a positive mammogram result has breast cancer) and provide better information to their patients if a premise containing statistical information is put in the form of 'Ten out of every 1,000 women have breast cancer' rather than 'The probability that a woman has breast cancer is 1% (prevalence)' (Gigerenzer *et al.* 2008: 55).

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

With this in mind, one possible way in which this knowledge could be used in a policy intervention is to change information about risk into a frequency format rather than a percentage or probability format which is less readily understood. That is to say that we change the environmental structure so that it is isomorphic with the decision-maker (in this case, it is assumed that the cognitive algorithms that underlie statistical reasoning are better suited to frequency formats). Alternatively, we could give people training in conversion between either sort of representational format. That way, they will be able to consider risky decisions using both sorts of information and come to a more informed decision. That kind of boost intervention would be construed as an extension of the heuristic repertoire of the targeted individual, which allows them to make their own alterations of the relevant environmental structure (Grune-Yanoff and Hertwig 2016: 156–159; Hertwig and Grune-Yanoff 2017: 977).

The contrast we have just given is suggestive of the way in which we might apply a type-distinction approach to the normative problems regarding BPP. First, the example of the framing nudge appears to bypass rational deliberation in a way that is morally ambiguous, and which gives rise to claims that it is manipulative or insufficiently respectful of autonomy. Things are quite different in the boost example, since either: (i) the policymaker has just restructured the environment so that it is better suited to the capacities of the individual (plausibly, a de-biasing of the choice architecture); or (ii) the policymaker has added to the heuristic repertoire of the individual. In either of these cases, it does not appear that the policymaker bypasses rational deliberation in the same morally ambiguous way. That suggests that boosts are respectful of autonomy in a way that nudges possibly are not, which would mean that we can determine that a BPP intervention is liberal in virtue of its being a boost rather than a nudge.

This suggestion is hinted at in Grüne-Yanoff and Hertwig (2016) and developed further in Hertwig and Grüne-Yanoff (2017). They say that boosts mark a normative difference from nudges because they require motivation and active cooperation on the part of the individual. That means both that they require transparency and that they respect autonomy.

> Boosts, in comparison [to nudges], require the individual's active cooperation. They therefore need to be explicit, visible, and transparent. The requirement of cooperation also implies individual judgment and engagement. This, in turn, implies – according to dominant notions of autonomy – that boosts are more respectful of autonomy than nudges are. (Hertwig and Grüne-Yanoff 2017: 982)

For example, in teaching consumers of medical information (e.g. patients or potential patients) how to convert statistical information into a

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

frequency format, they make no assumptions about the preferences of those consumers; they just boost their competences so that they can better satisfy whatever those preferences are.

What would be required in order to make this general line of thought into a viable framework? What is required is to distinguish the two sorts of policy intervention on grounds that are independent of the normatively loaded features at issue. We cannot just say: boosts are those policy interventions that, by definition, require transparency and which assume the autonomy of the individual; it is these policy interventions which are not vulnerable to normative critique. The first reason we cannot say this is that there is no guarantee that all boosting interventions do in fact require transparency and assume autonomy. We note that this is a claim in contradiction with what Grüne-Yanoff and Hertwig (2016: 176) themselves avow: 'boosts require minimal competences and motivations on the part of the target audience to be effective'. However, we don't see that this is obviously true. For example, consider a policy intervention which involves altering the environment of decision-making so that it is better suited to the individual's heuristics. This is a plausible example of a boost, because it is based on an ecological interpretation of bounded rationality. But it does not involve any need for motivation or cooperation on the part of the individual. We will also show that a similar ambiguity exists in cases where the heuristic repertoire is expanded (Section 3.1).

We might instead build these requirements into the definition. In that case, the instance just mentioned is not a boost because it fails to meet those requirements. So, all boosts are necessarily transparent and autonomy-preserving. But then, we are simply assuming what needs to be proved, and the argument is circular: Which BPP interventions respect autonomy? Boosts. And how are boosts distinguished from nudges? Boosts are different from nudges in virtue of their transparency and autonomy-preservation. It therefore appears that if the type-distinction strategy is to work, we need to establish two things on independent grounds: (i) the distinguishing features of boosts in contrast to nudges; and (ii) the features of boosts which make them immune from normative critique. The latter may be consequences of the former (for example, it might be that boosting a person's heuristic repertoire necessarily requires their cooperation, and that cooperation necessitates transparency and respect for autonomy), but they cannot be built into the original definition on the pain of circularity.

3. TWO STRATEGIES FOR MAKING THE DISTINCTION

The distinction between nudge and boost needs to be rigorous in just this way, if it is to do the work that we want it to do in marking a substantive normative difference. If it is not – if it admits of exceptions, or vagueness – then we can't be sure that it will exclude all illiberal BPP interventions from the boost category; that makes the type-distinction approach of little use to us. The reader will see that we are assuming a rather strict notion of classification in this case, where it is a determinate matter of fact as to whether a given BPP intervention is a boost or a nudge (or something else). To see why, imagine that you are a policymaker needing to determine whether a proposed intervention is in line with the requisite liberal values. In order to apply the type-distinction strategy you only need to know one thing: what category the token falls into. Such membership entails whether or not the token is liberal. But what if type-membership is graded or vague? Will the proposed entailment hold?

It's our view that graded membership won't do in this case because then in any case the entailment will only hold at certain grades, which makes the classification in effect as strict as we suppose above. And for vague membership, it also seems to us that such an entailment cannot hold either, because whether or not an intervention is transparent and respectful of autonomy is binary: either an intervention is understood by the individual or it is not; either an intervention preserves the individual's ability to choose or it does not.

To this end, the category of boosts needs to define a class of BPP interventions in contrast to two nearby categories: *nudges* and *non-BPP interventions*: these latter being all sorts of policy interventions that don't share the distinctive features of BPP. These would include traditional sorts of efforts to educate or provide information to a population, as well as more direct interventions that change the incentives on choice by applying rewards and punishments on permissible choices, as well as outright coercions, prohibitions and prescriptions. There are two reasons why boosts need to be distinguished from non-BPP interventions as well as nudges.[3]

The first reason, in short, is that the category of non-BPP interventions includes illiberal policy interventions. We take this to be obviously true: there is the potential for illiberal consequences (like manipulation or lack of respect for autonomy) in evidently coercive interventions (e.g.

[3] An anonymous reviewer has suggested that anything can count as BPP unless one has an uninterestingly narrow definition of BPP, one which is probably not shared by the authors we criticize. We demur from this judgement, on the grounds that the authors themselves call for a rather circumscribed understanding of BPP – for example, when they criticize others for offering *too* broad a definition: 'Thaler and Sunstein ... offered a sweeping compilation of supposed illustrations of nudges, including such general constructs as 'social influence' and 'social pressures'. Enlisting these social factors as drivers of behaviour change is, of course, nothing new' (Grüne-Yanoff and Hertwig 2016: 152–153). But even if we are mistaken in this assessment, we believe we have shown that these authors *must* draw the distinction in a narrower way if it is to supply a way to distinguish permissible from impermissible interventions.

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

punitively high taxes attached to cigarettes or alcohol) as well as milder forms of education and the provision of information which may be covertly illiberal. As an example of the latter, consider 'abstinence-only' sexual education in some parts of the United States and in developing countries, which arguably inculcates populations into a particular moral outlook. If the boost category cannot exclude non-BPP interventions, then it risks including illiberal policy interventions. That would make the type-distinction strategy defunct; we would then need to fall back on evaluating interventions on a case-by-case basis in independence of their type.

The second reason to distinguish boosts from non-BPP interventions is that we may otherwise occlude what is new, interesting, and distinctive about BPP in general. BPPs are all of these because they represent a new *evidence-based* approach to policymaking that draws upon a vision of the human reasoner as boundedly rational; this is the reason why they are distinguished as a novel category. It uses this evidence in order to make targeted interventions into the human cognitive system, where this includes the context in which decisions are made. Indeed, this is why it is deeply unsatisfying when authors like Sunstein (2015) cast the category so wide as to include state interventions like a Bill of Rights; it effaces the novelty of BPP as a type of its own, and a genuine advance in evidence-based policymaking. Why should we be interested at all, if this is just old wine in new bottles?

With this in mind, the rest of this section discusses the ways in which the distinction could be made rigorous enough to mark a genuine normative difference, and so underwrite the type-distinction approach to liberal BPP. We consider two different strategies for making this distinction.

1. **To define the category in an intuitive manner through the use of paradigmatic examples and distinctive properties.** Certainly, choosing the right examples of either kind can make the difference appear pretty stark. The cases of energy conservation and frequency formats demonstrate different approaches to policymaking. In one, it seems, the decision is framed so that the biased decision-maker is more likely to choose the option that they would evaluate more highly under ideal conditions. In the other, the decision is framed so that the individual's heuristics are better suited to the environment in which the choice is made. It may be possible to draw the distinction in a paradigmatic fashion using a collection of such examples. In doing so, we may extract a number of distinctive properties which help us distinguish new cases as either boost or nudge interventions.

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

2. **To draw the distinction by means of the causal mechanisms by which each operate.** This means specifying the causal variables that are manipulated by either sort of policy intervention, and perhaps the different ways in which those interventions manipulate those variables. For example, it may be that boosts causally intervene in an individual's competences, rather than on her behaviour. A reasonably full specification of contrasting mechanisms of action would suffice to distinguish the two policymaking approaches.

We consider these two strategies in turn.

3.1. Paradigmatic Examples and Distinctive Properties

Can a handful of the right examples help us distinguish boosts from nudges in a way that can determine which interventions are permissible? Here's how that approach might work:

1. Assemble a collection of paradigmatic examples for either type.
2. Extract a list of properties or assumptions which set the examples apart.
3. Show that one or more of these properties or assumptions make the type immune to the relevant normative criticism.
4. Apply the properties or assumptions in classifying BPP interventions on a case-by-case basis, allowing us to classify them both descriptively (in virtue of step 2) and normatively (in virtue of step 3).

We think that this strategy will fail at step 3, and possibly also at step 4. It fails at step 3 because, *contra* Hertwig and Grüne-Yanoff (2017), none of the paradigmatic traits attributed to boosts (for example, cooperation and motivation on the part of the individual) are able to entail the required respect for autonomy or transparency. It may also fail at step 4 because the examples which are supposed to be paradigmatic of either kind are strongly underdetermined in the absence of a more rigorous theory of the distinction that appeals to underlying causal mechanisms. That is to say that the mere behavioural details of each case can be interpreted in multiple ways. But let us examine the strategy in more detail, as it is central to both of the papers that we criticize.

We will consider three paradigmatic examples of boost interventions which boost the heuristic repertoire of the targeted individual, ignoring policies which only alter the environment in order to achieve short-term gains (for example, an immediate but temporary improvement in the statistical reasoning of the individual). That is because we think that the best case for demarcating boosts as a distinct category of BPP

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

lies with these, what Hertwig and Grüne-Yanoff (2017) call 'long-term boosts'. Indeed, the authors themselves pick them out as a category most obviously distinct from nudges. Long-term boosts are intended to produce changes in behaviour that survive beyond the intervention itself, and do so by boosting competences in a way informed by behavioural science.

3.2. Steps 1 and 2: Paradigmatic examples and their properties

3.2.1. Example 1: Heuristics to improve financial outcomes. Hertwig and Grüne-Yanoff (2017: 978–979) introduce a hypothetical intervention that is intended to be a boosting counterpart to Save More Tomorrow. This boost is intended to have people save money, in response to the same problem: people tend not to save enough money for retirement, even if they wish to. They model this hypothetical intervention on an accounting intervention which trains individuals to keep business and personal accounts separate through keeping the money in two separate drawers, and only transferring money along with an IOU note that tracks the transfer. That allows them to keep track of profits without any sort of theoretical knowledge from accounting. The suggestion is that this replacement of explicit knowledge with practical heuristics could take place in the domain of saving and investment, as well (they suggest 'a simple 1/N diversification strategy' (cf. DeMiguel *et al.* 2007); we might also suggest investment in index funds rather than in specific companies).

They also suggest that policymakers could boost the competence of individuals to feel a sense of continuity over time, so that they feel that to save money is not to be sacrificing it. To do this, they suggest computer programs that predict the way the individual might look at some point in the future by artificially ageing an image. Individuals who underwent this procedure were more likely to exhibit financial self-control in trials (Hershfield *et al.* 2011). Training people to feel a sense of psychological connectedness over time through such procedures might produce a boosted competence to choose larger rewards that are more distant in time against short rewards that are more immediate – that is to say, to temporally discount reward in a more rational fashion.

3.2.2. Example 2: Decision trees for diagnosing clinical depression. One sort of boost intervention aims to equip individuals with simple heuristics which aid decision-making in complex domains. One example of this is medical decision-making. Grüne-Yanoff and Hertwig (2016: 160) give the example of the diagnosis of depression, which can be a time-consuming process when using the standard 21-item questionnaire. They cite a study by Jenny *et al.* (2013) which used a fast and frugal decision tree enabling physicians to diagnose depression much more quickly and up to a high level of accuracy. This decision tree contains just four questions which are

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

asked in sequence, and for which each question except the last terminates either in a negative diagnosis or in the next question. The final question diagnoses either in a positive or a negative diagnosis. This heuristic method was found to perform better than competitor techniques (except for the standard method) of greater complexity.

3.2.3. Example 3: Training in the conversion of statistical information to different representational formats. This policy intervention was briefly discussed in Section 2. In the conclusion to their study on representational formats and statistical reasoning, Gigerenzer and Hoffrage (1995) suggest that the right approach to systematic error in this area is to teach 'representations' rather than 'rules'. What they mean by this is that an effective approach to improving people's statistical reasoning would be teaching them to convert probabilities into frequencies, rather than attempting to teach them more complex rules like that of Bayes' theorem (Sedlmeier and Gigerenzer 2001). This is an instance of a more general sort of boosting strategy: 'brief training in transforming opaque representations (e.g., single-event probabilities) into transparent ones (e.g., frequency-based representations)' (Hertwig and Grüne-Yanoff 2017: 979). The idea behind this intervention is that our competences are better suited to frequency formats, and that we can intuitively deal with statistical problems involving that format without needing to apply the complex rules of probability theory.

From these examples and others like these, one may identify a number of distinguishing features, as indeed Hertwig and Grüne-Yanoff (2017: 974) do. First, it appears that (long-term) boosts target the competences of the individual rather than the behaviour of the individual. The difference here is that behaviour can be changed without making any alterations in the competences of the individual, whereas to change competences is also to change behaviour through changing an important proximal cause of that behaviour. It is therefore likely to continue to have effects on behaviour when the successful intervention is removed. Furthermore, depending on how general the boosted competence is, it may also have beneficial effects over multiple contexts. Last, and most importantly, to boost the competences of the individual in this way we can see that this requires a minimal *cooperation* on the part of the individual as well as a minimal *motivation* to use the acquired competence or heuristic in the appropriate situations.[4]

[4] Indeed, Hertwig and Grüne-Yanoff (2017) also consider causal/mechanistic criteria: the malleability of cognitive architecture and causal targeting of heuristics/decision-tools. We consider these criteria in Section 4, because they are deeper and more likely to overcome the underdetermination of the examples.

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

3.3. Step 3: No Descriptive Assumptions Entail Normative Immunity

Our first claim will be that none of the properties extracted from the policy instances above entail immunity from the same normative concerns that attach to nudging policy interventions. To see this, consider those properties that the authors specifically single out as entailing transparency and respect for autonomy. These are that it requires the cooperation and motivation of the targeted individual to work, and that it is therefore by necessity 'explicit, visible, and transparent' (Hertwig and Grüne-Yanoff 2017: 982).

We can see this property in all three of the examples considered above. In all of these cases, the intervention expands the heuristic repertoire of the targeted individual but it requires the cooperation of the individual in order to do so: this seems true because (i) cooperation is required to learn the heuristic in the first place, and (ii) the agent needs to be motivated in order to use the newly learned heuristic at all.

Why, however, should we think that active cooperation on the part of the individual necessitates transparency and respect for autonomy? We don't think this is obvious. To see why not, we can consider cases of policy interventions which satisfy this cooperation requirement but which do not seem to be transparent or respectful of autonomy. Consider again the case of abstinence-only sexual education.[5] This is a policy intervention which requires the cooperation of the targeted individual.[6] It assumes a goal in a particular context: either the desire to avoid STIs or unwanted pregnancy. Then the targets of the intervention, if their preferences conform with this goal, can choose to engage with the intervention. The intervention is simple, and adds to the heuristic repertoire of the individual. That heuristic is just this: don't have sex before marriage. This is a simple heuristic which has a guaranteed chance of success if it is widely practiced, both in reducing individual risk and in improving any associated epidemiological factors (e.g. spread of disease). Some abstinence-only programmes also draw on behavioural science in order to help the targeted individuals, drawing upon resources from 'social cognitive theory, social inoculation … the health belief model, and cognitive behavioural theory' (Underhill *et al*. 2007).

[5] We may appear inconsistent in talking about this as a boost, since earlier on in this paper we refer to abstinence-only sexual education as a non-BPP. But this is not inconsistent, because we don't believe that the distinctive properties considered in Section 3.1 are sufficient to distinguish boosts from non-BPP.

[6] Arguably this cooperation is sometimes implicitly coerced, if it is given in the context of a school curriculum where cooperation is mandatory. But this need not be so – if, for example, when it is applied as an attempt to reduce the spread of HIV in developing countries (e.g. Uganda, cf. Stoneburner and Low-Beer 2004).

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

This intervention has the distinguishing properties of a boost-type intervention that are putatively relevant to its normative status. It targets competences, fosters these competences through changes in skills and tools for decision-making, assumes cognitive malleability, aims for persistent effects across various contexts, and requires the individual's cooperation in the programme. However, abstinence-only sexual education programmes have come under heavy criticism for being illiberal (Santelli *et al.* 2006) and heteronormative (Wilkerson 2013), as they inculcate a specific moral outlook. So it is not at all obvious that the need for cooperation entails transparency and respect for autonomy.

One possible reply to this counterexample is that in fact abstinence-only programmes, to the extent that they are illiberal, are so just because they lack the required transparency. That is to say that the interventions do not adequately teach about alternative methods of preventing unwanted consequences of casual sex. They are not transparent about these alternatives. That means they are not properly considered boosts. But this reply is not very convincing: all it does is show that transparency can't be read off of the fact that cooperation is required on the part of the targeted individuals. And as we argued in Section 2, to include transparency directly in the definition of boosts rules out the use of the distinction to help distinguish permissible from impermissible policy interventions – such a move would be unacceptably circular.

3.4. Step 4: Individual examples are underdetermined

For the sake of argument, however, let's temporarily grant that Step 3 could be successful, so that some of the descriptive requirements of boost policy interventions (along with some reasonable normative assumptions) entail their permissibility. After that, Step 4 of the procedure outlined above requires that we apply the descriptive features in order to classify new instances. Assuming the success of Step 3, the classification of an instance into the boost category would then allow us to read its permissibility off from its type-membership. However, we think that this step won't be possible without a deeper causal story about what is going on in either policy type. Establishing the truth of this claim will allow us to go on to consider various candidates for that deeper causal account. The way we propose to do this is to show how the examples that we've considered can be interpreted either as (i) a nudge intervention; or (ii) a non-BPP intervention. The only way to decide between these competing interpretations is by the principled application of a causal model which specifies how the two policy approaches differ in their mechanism of action.

Consider first the financial boost of Example 1. This example contained two distinct strategies, or ways in which the competences of the

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

individual might be boosted in order to change saving behaviour. The first was to teach simple heuristics which do not require theoretical knowledge about finance but which will reliably result in higher levels of savings or return on investment – for instance, 'a simple 1/N diversification strategy' where wealth is spread out evenly amongst a number of different investment assets. DeMiguel *et al.* (2007) show that such a strategy performs just as well as more complex strategies, and it is clear that it is easier for most investors to apply.

However, the interpretation of this intervention as a boost is not unambiguous. Use of rules of thumb, even if they appear to have beneficial effects in certain experimental contexts, are not appropriately evidence-based in the absence of a mechanistic model of how they work to intervene on the cognitive or motivational structure of the target (Grüne-Yanoff 2016), and should not be considered BPP, which is an evidence-based approach to policymaking that aims to directly make interventions in the competences of the decision-maker on the basis of demonstrated knowledge about bounded rationality.

The second financial boosting strategy was to show individuals how they might look at some point in the distant future, in order to boost psychological connectedness and aid their competence in offsetting the overvaluation of immediate rewards. However, it's not clear why this should be considered a boost rather than a nudge. That is because one can see this intervention as a framing effect that switches from a short-term frame to a long-term or team frame (Gold 2013), and that is an intervention that acts without necessarily boosting the competence of the individual; it just allows them to see the problem in a new frame. So interpreting this intervention as a nudge is consistent with the improvements that individuals show in self-control after the intervention.

Consider Example 2. The example concerns the use of fast and frugal decision trees in order to make decisions in the medical domain – in this case, for the diagnosis of depression. But it's not so clear that this deserves to be classified as a boost, in distinction from both nudging and non-BPP interventions. The reason for this is again because teaching simple rules of thumb is not necessarily a BPP intervention. To teach people such rules does not require any evidence base from behavioural science at all. What would be needed in order to show that this is a BPP is a deeper story about the mechanisms or causal variables that are at play in the specific decision-problem (here, diagnosis), and a demonstration about how these mechanisms or variables are intervened upon in a way that differs from a nudge intervention, and from mere information provision.

The last paradigmatic example – the conversion of statistical information between different formats – seems to us the strongest and most apparently unambiguous of all the examples. However, this clarity is only apparent. That is because these two forms of statistical representation

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

are not equivalent in the information they provide, even though the example presents what is going on as a mere de-biasing. That would also mean that to teach people to switch formats is not a true boosting of competences; it may sometimes result in poorer or more confused decision-making. This is true both for the switch between frequency formats and probabilities, as well for the switch between absolute and relative risk. Rather, the forms of statistical representation that tend to make people better reasoners include more information; *they include base rates*. Take this instance that is cited in Grüne-Yanoff and Hertwig (2016: 158; originally from Gigerenzer *et al*. 2008):

> *Relative risk reduction*: if you have this test every 2 years, it will reduce your chance of dying from this cancer by around one third over the next 10 years. *Absolute risk reduction*: if you have this test every 2 years, it will reduce your chance of dying from this cancer from around 3 in 1000 to around 2 in 1000 over the next 10 years.

This could be considered a boost, on the following assumptions. First, we must assume that the presentation of the information in the absolute format constitutes a *de-biasing* of the choice architecture. If it is true that the cognitive algorithms that produce judgements to statistical problems are suited to frequency formats, then we might infer that changing the information into this format is a way of removing pressures leading to inaccurate judgement. Second, we must assume that more accurate statistical judgement on such occasions will make the individual more competent at decision-making *per se* on those occasions. That is to say that they are more likely to make the 'right' choice with respect to treatment, or finance, or some other domain.

However, these two assumptions are up for contention. First of all, it is not so clear that this is a mere de-biasing; the provision of additional information (base rates) and the consequent shift of focus will change our interpretation of the facts. It will be a different account of valuable information that may be just as (or more) biased as the other. For example, it may lead individuals to inappropriately underestimate their levels of risk if other factors (e.g. potential rewards or penalties) which make the threshold of risk lower are absent. For example, it may be more appropriate to take the test above if one has a history of that cancer in one's family, even if the reduction in risk is seen to be relatively small when viewed in absolute terms. Likewise (if the reduction of risk is high), it may be less appropriate to take it if the potential penalty for not doing so is negligible. Our point is that these extraneous factors play a role in appropriate judgement, making over- or under-estimation of risk appropriate in some circumstances.

Even so, we might emphasize the fact that the individual is taught how to convert between these two formats. That might mean that we

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

have taught them a new competence: they can switch back and forth between these formats, consulting their intuitions in each case, and then make a more informed decision under conditions of uncertainty. However, we *still* don't think it's obvious that this has boosted the competence of the individual by enhancing their heuristic repertoire. It has merely given them a way in which to acquire more information to consider during their deliberation on a particular decision concerning risk. More information is not always better, it must be *relevant* information: the intuitions produced in the individual when she considers relative and absolute risk will not necessarily be relevant unless the broader context of the decision is understood. This broader context will include epidemiological information as well as the rewards or penalties associated with the choice.[7]

To conclude this subsection, we submit that there is something going on here that is deeper than mere infelicity. Rather, it seems to us that these examples are underdetermined for two key reasons: (i) without a mechanistic theory of the distinction, there's no way to distinguish in a principled way between a genuine improvement in competences and mere behaviour change; and (ii) without a mechanistic theory, there's no way to draw a principled link between the policy intervention and the relevant behavioural science; the relationship will just be suggestive or heuristic, rather than genuinely evidential (cf. Grüne-Yanoff 2016). That is why putative examples of boost intervention are underdetermined, and can be reinterpreted as either nudge interventions or non-BPP interventions. For this reason, we think it is not worthwhile showing how every single paradigmatic example admits of ambiguity; one should rather look for a deeper approach which could eliminate this ambiguity.

4. DIFFERENT CAUSAL MECHANISMS FOR NUDGE AND BOOST

To make the distinction between nudge and boost clear and exceptionless in the way that would be required, we need a mechanistic theory that specifies the way in which each operates on different causal variables (perhaps in different ways). This view appears to be shared by the authors of the boost approach, who make some gestures in this direction. In their discussion of the under-determination of one of the examples they present, Grüne-Yanoff and Hertwig (2016: 163) say this:

[7] In his report to *Economics and Philosophy*, one reviewer suggests a particular situation for which this intervention constitutes a boost: when actors in the medical industry overemphasize the benefits of treatments for avaricious reasons. Then it boosts the competence of individuals to be savvy about misleading advertising. But then the individual has to make a choice about whether to purchase the treatment nonetheless, and the problems recur in this context.

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

> The SH [simple heuristics] and H&B [heuristics and biases] programs tend to disagree about the underlying *causal mechanism* that explains the relationship between default setting and changes in choice distribution. H&B researchers tend to explain default effects in terms of "inertia, status-quo bias, or the 'yeah, whatever' heuristic" … This kind of explanation stresses the biasing features of setting any default, thus revealing the policy to be *re*biasing. SH researchers, in contrast, explain default effects in terms of the implicit recommendation or endorsement effect … This kind of explanation stresses the genuine social information contained in the default, thus describing the behavioral change in response to the default as consisting in a learning effect, and hence revealing the policy to be *de*biasing. Thus, even in cases where nudgers and boosters propose the same policy, their respective mechanistic interpretation of the intervention distinguishes the distinct goals they pursue with it.

In this section we discuss the prospects for this sort of strategy. Although this has not yet been worked out in detail, there are suggestive remarks in Hertwig and Grüne-Yanoff (2017) which indicate a few possible ways in which it could be pursued.

4.1. Dual-process Architecture Versus Cognitive Malleability

In Hertwig and Grüne-Yanoff (2017: 979–980), it is proposed that one distinguishing difference between nudge interventions and boost interventions is in the assumptions that policymakers of either persuasion make about cognitive architecture. Theories of cognitive architecture are theories about the overall structure and function of the human cognition, when this is construed both in terms of subpersonal information processing as well as the mental entities (beliefs, goals …) that supervene on that information processing. One way to understand this is in terms of the stable features of a cognitive system that remain invariant across changes in the informational content of the system (Pylyshyn 1984). In this context, nudging is thought to imply the assumption of a *dual-process* architecture (cf. Evans and Stanovich 2013) while boosting is thought to imply the assumption of 'cognitive malleability'.

Dual-process theories of cognitive architecture state that there are two cognitive systems (or two distinct types of cognitive process), where one of these types ('System 1') is fast, intuitive, and stereotyped in its operations; and where the other ('System 2') is slow, deliberate, rule-governed, and places a heavy load on working memory. Nudging – at least for the 'bias-recruiting' nudges which are central to the normative controversies – is thought to operate on System 1 processes in order to nudge individuals into better behaviours, in virtue of the fact that their System 2 processes are not efficacious enough for many of them to do so alone.

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

This view is at odds with the assumption that cognitive architecture is cognitively malleable, a view that 'proponents of boosting necessarily agree on' (Hertwig and Grüne-Yanoff 2017: 980). The views are at odds because the heuristics which are acted on by nudging policymakers are assumed to be stereotyped and incapable of change – they are to some extent 'fixed' in their operations, except when the individual employs System 2 resources to tackle the same decision problem. For boosting policymakers, on the other hand, it is assumed that the heuristics employed by individuals are malleable in two ways: (i) that they form an 'adaptive toolbox' that can be augmented by the right kinds of interventions; and (ii) that the competences constituted by this toolbox can be further enhanced by re-structuring the environment (or teaching individuals how to do this themselves) so that it is isomorphic with the relevant heuristics.[8]

These two sets of assumptions are not just different but in fact incompatible. The malleability of heuristics that is central to the boost approach is denied within the dual-process framework, for which these heuristics are automatically triggered in the right circumstances in a way that is autonomous from reflective processes, and despite the sort of format in which the relevant information is given. The difference can be seen more clearly in the divergent ways that Gigerenzer and Hoffrage (1995) and Kahneman and Tversky (1996) interpret the discovery that frequency formats can improve statistical reasoning in naïve subjects. The latter authors do not deny that changing the format improves performance, but they do deny that this is a genuine change in System 1 heuristics which they argue remain fixed in their effect as 'cognitive illusions'. Rather, the improvement in performance can be explained as the introduction of additional informational cues which enable individuals to avoid error. They point out that similar systemic errors caused by those heuristics persist even in cases which include information about frequencies, and suggest that this is further evidence for the fixity of System 1 heuristics.

But if the assumptions are incompatible, then to distinguish nudge from boost interventions in this way is counterproductive. That is because neither nudgers or boosters will be in a position to say that the policy types are distinct (at least not on the basis of distinct assumptions about cognitive architecture); insisting on cognitive architecture as a difference means that they will rather be committed to assimilating the other approach under their own. For example, boosters will be committed to saying that nudges are just a sort of 'short-term boost', about which the proponents of nudging are mistaken with respect to the deeper causal processes. Likewise, nudgers will be committed to saying that boosts are nothing more than a sort of 'educative nudge', about which the boosters

[8] Hertwig and Grüne-Yanoff (2017) note that boost interventions are not strictly limited to cognitive heuristics, but may potentially include interventions on motivation.

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

are mistaken vis-à-vis the causal processes. So this distinction between cognitive architectures is not appropriate as a way of eliminating the ambiguity of the particular policy instances.

4.2. Intervening on Behaviours Versus Intervening on Competences

This is a suggestion about the causal variable that is acted upon by either kind of policy intervention. It is weaker than the architectural approach and so more likely to be congenial to pluralism about BPP. On this approach, it is assumed that there is just one cognitive architecture in a decision-making context – the nature of this cognitive architecture does not necessarily need to be specified in detail – and that there are different causal variables within this architecture or context that are acted upon by distinct types of BPP. In the case of nudging and boosting interventions, these are that nudges act in order to change individuals' behaviours while boosts act in order to change individuals' competences.

Take the contrast between Save More Tomorrow and a hypothetical savings boost. We may distinguish these in terms of the causes being intervened on: Save More Tomorrow induces behaviour change through clever alterations in the choice-architecture, while a savings boost enhances the person's competence in saving by, for example, fostering psychological connectedness with the person's future self through artificial ageing (Hertig and Grune-Yanoff 2017: 978; see Section 3.1, 'Example 1').

But now, how do the authors propose that we distinguish competence-change from mere behaviour-change? Why, for example, can't we see the Save More Tomorrow intervention as boosting the competence of the agent in their ability to value rewards by altering the structure of the environment so that it is better suited to the way that individuals discount future reward? That would be perfectly consistent with the way in which Hertwig and Grüne-Yanoff (2017: 980) characterize their architectural commitments: 'competences are often best fostered by redesigning aspects of individuals' external environment or by teaching them how to redesign them'. Likewise, why can't we see the 'artificial ageing' as a nudge that elicits a 'team frame' in the individual so that they temporally discount in a manner that is closer to optimal? That would not require any change in the competences of that individual.

The reply to this point may be that we are missing the role that the agent plays in this distinction: for boosting, what is important is that the agent may apply or ignore the taught heuristic as she sees fit. That is not the case in Save More Tomorrow, but it may be if it was designed so that the agent was taught to switch between one or another temporal perspective at will. The problem with this way of salvaging the 'intervening on competences' criterion is that it is not really a causal or mechanistic criterion at all. This suggestion would rather be

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

a pragmatic or behavioural guideline associated with nudging policies, and that would putatively make them ethically sound: when applying nudge policy interventions, teach the target of the intervention to apply this themselves. We discussed such a guideline in Section 3.1. The problem is that: (i) this is not possible for all cases, as in the case of the default option; and (ii) this does not in any way guarantee that the policy will be immune from normative critique (Section 3.1, 'Step 3').

In short, the problem we see here is that you need a framework in order to distinguish changes in competences which entail changes in behaviours from mere changes in behaviours that don't imply changes in competences. To say that these are different causal factors which are intervened upon will not be adequate unless there are criteria allowing us to identify those factors in particular cases. This is not forthcoming, and so the ambiguity remains in the cases. However, the authors do suggest a more pragmatic and empirical approach to determining the cases in which competences have been enhanced.

4.4. The Reversibility Criterion

Related to these mechanistic strategy is a more pragmatic proposal that *would* work in a way compatible with pluralism. Specifically, Hertwig and Grüne-Yanoff (2017) suggest an empirical criterion by which to distinguish boosts from nudges, by allowing one to distinguish behaviour-change from competence-change in a reliable manner:

> If, *ceteris paribus*, the choice architect eliminates an efficacious (non-monetary and non-regulatory) behavioural intervention and behaviour reverts to its pre-intervention state, then the policy is likely to be a nudge. If, *ceteris paribus*, behaviour persists when an intervention is eliminated, then the policy is more likely to be a boost. (Hertwig and Grüne-Yanoff 2017: 981)

The idea behind this pragmatic criterion is related to the mechanistic strategy for distinguishing boosts from nudges. More particularly, it is based on the conjecture that boosts produces behaviour change through expansion and enhancement of agent competences, where nudges do the same by bypassing those competences, leaving them unchanged. Since the proximal causes of the changed behaviour are internal to the agent, on the mechanistic analysis of boost, those causes continue to operate in the absence of the intervention. The proximal cause of behavioural change for nudges are changes in the choice architecture which recruit agents' biases, and so should only operate while that intervention is in force.

There are two points to be made with respect to the reversibility criterion in the context of our discussion. The first is that the criterion as thus stated is clearly loose and intended only as a practical heuristic; the *ceteris paribus* clauses and other hedging devices in the criterion statement

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

make that clear. With this in mind, it seems unlikely that the criterion will be able to exclude nudges and non-BPPs in the way that is required to apply a type-distinction strategy to the ethics of BPP. Here's a puzzle to do with the criterion that demonstrates this point: Hertwig and Grüne-Yanoff (2017) countenance that nudge interventions may have lasting effects. For example, being in an intervened-on choice architecture for a long period of time may form habits in the individual – a tendency to prefer healthy food, for example. They say that in these cases the intervention is not to be construed as a nudge but rather that the nudge has a 'boosting side effect': 'The nudge has thus turned into a boost and had lasting effects' (p. 9). The difficulty here is that such a side effect will evince properties that Hertwig and Grüne-Yanoff take to be *un*characteristic of boosts. According to them, boosts always involve a collaboration between the policymaker and the individual, such that they require active cooperation on the part of that individual. But habits learned as a result of covert manipulation of choice architecture are clearly not the result of collaboration of this kind.

The second point is that the criterion is limited in its use in the context of normative evaluation. It is limited because it can only be applied *after* the intervention has been put in place. When we are evaluating policies for their liberality, we want an answer to this prior to their implementation. If we only have a post-hoc criterion, such as the reversibility criterion, then we must implement a policy prior to knowing whether it is permissible by the lights of our background ethical guidelines. That means potentially violating those guidelines; implementing an illiberal policy, for instance, and violating citizen autonomy. It is therefore preferable, if possible, to implement an approach to evaluation that is not post-hoc in this way.

With these points in mind, we suggest that the reversibility criterion – while undoubtedly an innovative and useful rule-of-thumb – is not sufficient in the context of applying a type-distinction strategy to the ethics of BPP.

5. LET'S KEEP PERMISSIBILITY OUT OF IT: TYPOLOGY AND VALUE

It's on the basis of the considerations above that we suggest that the type-distinction strategy does not work in resolving the normative questions concerning BPP. We hope that the reader does not take this to mean that we are endorsing a rejection of the nudge/boost distinction entirely; we think that any such rejection would be impractical and unwise. Rather, we just think that the distinction cannot do the right kind of work in being the basis for the approach to answering normative questions in the study of BPP. And in fact, we think that the distinction between nudge and boost has great heuristic value in suggesting new interventions to policymakers, as well as expanding the scope of BPP beyond the narrow view of bounded rationality implied in the Heuristics and Biases

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

research programme. So we agree that Grüne-Yanoff and Hertwig are to be commended for a significant contribution.

How, then, *should* we attempt to answer the normative question? Our view is just that policy interventions need to be evaluated on a case-by-case basis, and not by reference to some subtype under which they are categorized (like nudge, or boost). On the type-distinction strategy, knowing that something is a boost (say) is enough to know that it respects autonomy, because in the way they are defined boost interventions make no assumptions about the preferences of the individual. We think that this approach is far too complacent: respect for autonomy cannot just be read off from type membership in this way. But how might case-by-case evaluation of BPP be performed?

One possibility would be to argue for the special relevance of a particular theory of autonomy from the philosophical literature (Buss 2016), and to use this as a yardstick against which to evaluate particular interventions. One way to avoid complications resulting from philosophical disagreements might be to synthesize these in order to produce a theory of autonomy which is silent on the central points of controversy, but which supplies a set of minimal criteria on which all parties to these debates could agree, and that may potentially be breached by BPP interventions.

Another would be to first produce a normative evaluation of the heuristics and biases on which BPP interventions rely. For example it may determine, for each heuristic, whether and how that heuristic has the potential to undermine autonomy. For example, there seems to be a rather clear difference in the way that the representativeness heuristic is related to some particular rule-of-thumb: given the automaticity of the representativeness heuristic, it cannot be applied in an autonomous manner. However, some learned rule-of-thumb has the potential to be so applied. Furthermore, presumably 'autonomy' requires as a necessary condition the good function of various cognitive abilities that allow individuals to weigh up their values, discriminate between various possibilities, and so on. It could therefore be that a better empirical understanding of these abilities themselves provide the basis for a framework for normative evaluation. This may be combined with the evaluation of biases: such a study would determine to what extent those biases undermine these autonomy-making capacities in the context of a particular policy intervention.

This approach that we sketch is necessarily incomplete given the scope of the paper, but we can already see some significant differences from a type-distinction approach. The first is that the type-membership of the intervention is irrelevant to its normative evaluation. That has the welcome consequence of distinguishing normative from descriptive characterizations of some particular policy intervention. The second

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

is that the social context of the intervention may play a larger role in its evaluation. We believe that some policy interventions may be permissible in some contexts but not in others. For example, it may be that behaviourally informed abstinence-only sexual education may be permissible in the context of adult education but not in the context of school education (even if non-compulsory), since adults are already developed in their moral outlook. To read off permissibility from type-membership effaces these important contextual factors, since those contexts are variant across and within policy types.

Finally, we suggest that liberating discussion of BPP typology from being mired in the normative concerns will afford a clearer view on what sorts of types there really are. We can see this with respect to boosts: as Grüne-Yanoff and Hertwig describe it, a boost takes the form either of providing the individual with new competences or structuring the choice-architecture in line with existing competences. As an example of the former, recall the sorts of fast and frugal decision trees that were provided in order to quickly and reliably diagnose clinical depression. For the latter, imagine a kind of general policy to provide information about risk in absolute rather than in relative formats.

What do these two sorts of strategies have in common? We submit that the primary reason to call each of these 'boosts' is the general idea that they both bear this putative normative difference in opposition to nudge interventions. However, once we put those issues aside we can in fact see that they are quite different in character. It seems to us, then, that Grüne-Yanoff and Hertwig have discovered the possibility of not just one but (at least) two distinct types of BPP intervention. It may be that a fuller causal model – perhaps, for instance, one based on diffusion-to-bound models in computational neuroscience (Felsen and Reiner 2015) – than we currently have can help us distinguish these types in terms of the causal variables that they act upon within that model.

Whatever the approach, we are quite sure that the right way forward will be to produce an evaluative framework that is applied to interventions on their own terms, and not in the framework of type-membership. Such a type-membership approach is simply incapable of delivering reliable normative verdicts on individual policy interventions. Likewise, treating permissibility as a distinct issue will allow us to get a clearer view with respect to the typological questions.

ACKNOWLEDGEMENTS

Andrew Sims gratefully acknowledges funding from the Action de Recherche Concertée 'Free Will and Causality', and Thomas M. Müller would like to acknowledge funding from the INSOSCI interdisciplinary research group on 'Integration of Cross-Disciplinary Research in

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

Neuroscience and Social Science'. Both authors would like to thank an audience at the EIPE 20th Anniversary Conference and other members of the INSOSCI working group and the LED group for comments on earlier iterations of the argument.

REFERENCES

Barton, A. and T. Grüne-Yanoff. 2015. From libertarian paternalism to nudging – and beyond. *Review of Philosophy and Psychology* 6: 341–359.

Buss, S. 2016. Personal autonomy. In *Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. <https://plato.stanford.edu/archives/win2016/entries/personal-autonomy/>.

DeMiguel, V., L. Garlappi and R. Uppal. 2007. Optimal versus naïve diversification: how inefficient is the 1/*N* portfolio strategy? *Review of Financial Studies* 22: 1915–1953.

Evans, J. S. and K. E. Stanovich. 2013. Dual-process theories of higher cognition: advancing the debate. *Perspectives on Psychological Science* 8: 223–241.

Felsen, G. and P. B. Reiner. 2015. What can neuroscience contribute to the debate over nudging? *Review of Philosophy and Psychology* 6: 469–479.

Gigerenzer, G., W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz and S. Woloshin. 2008. Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest* 8: 53–96.

Gigerenzer, G. and U. Hoffrage. 1995. How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review* 102: 684–704.

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

Gold, N. 2013. Team reasoning, framing, and self-control: an Aristotelian account. In *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience*, ed. N. Levy, 48–66. Oxford: Oxford University Press.

Grüne-Yanoff, T. 2012. Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare* 38: 635–645.

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

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

Hershfield, H. E., D. G. Goldstein, W. F. Sharpe, J. Fox, L. Yeykelis, L. L. Carstensen and J. N. Bailenson. 2011. Increasing saving behaviours through age-progressed renderings of the future self. *Journal of Marketing Research* 48: S23–S37.

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

Jenny, M. A., T. Pachur, S. L. Williams, E. Becker and J. Margraf. 2013. Simple rules for detecting depression. *Journal of Applied Research in Memory and Cognition* 2: 149–157.

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

Kahneman, D. and A. Tversky. 1996. On the reality of cognitive illusions. *Psychological Review* 103: 582–591.

Pylyshyn, Z. W. 1984. *Computation and Cognition: Towards a Foundation for Cognitive Science*. Cambridge, MA: MIT Press.

Santelli, J., M. A. Ott, M. Lyon, J. Rogers, D. Summers and R. Schleifer. 2006. Abstinence and abstinence-only education: a review of U.S. policies and programs. *Journal of Adolescent Health* 38: 72–81.

Sedlmeier, P. and G. Gigerenzer. 2001. Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology* 130: 380–400.

Simon, H. A. 1956. Rational choice and the structure of the environment. *Psychological Review* 63: 129–138.

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press

Stoneburner, R. L. and D. Low-Beer. 2004. Population-level HIV declines and behavioural risk avoidance in Uganda. *Science* 304: 714–718.

Sunstein, C. R. 2015. Nudges, navigability, and abstraction: a reply to critics. *Review of Philosophy and Psychology* 6: 511–529.

Thaler, R. H. and S. Benartzi. 2004. Save More Tomorrow$^{\text{TM}}$: using behavioural economics to increase employee saving. *Journal of Political Economy* 112: S164–S187.

Thaler, R. H. and C. R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

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

Underhill, K., P. Montgomery and D. Operario. 2007. Sexual abstinence only programmes to prevent HIV infection in high income countries: systematic review. *British Medical Journal* 335: 248.

Wilkerson, A. 2013. I want to hold your hand: abstinence curricula, bioethics, and the silencing of desire. *Journal of Medical Humanities* 34: 101–108.

BIOGRAPHICAL INFORMATION

Andrew Sims is a post-doctoral Fellow at the Institut supérieur de philosophie at the Catholic University of Louvain (Louvain-la-Neuve). He is interested in the ways in which cognitive science can be used to inform social and policy interventions, particularly with respect to the public understanding of hyper-politicized issues in policy-relevant science. His current work also focuses on empirical challenges to classical models of human action and free will.

Thomas Michael Müller is an Assistant Professor (maître de conférences) in University Paris 8. He is an historian and philosopher of economics, and is interested in epistemological issues in philosophy of economics and philosophy of science. His current work also focuses on philosophy of neuroeconomics and behavioural economics, in close collaboration with the INSOSCI project (www.insosci.eu).

https://doi.org/10.1017/S0266267118000196 Published online by Cambridge University Press