35

BEHAVIORAL PUBLIC POLICY

One Name, Many Types. A Mechanistic Perspective

Till Grüne-Yanoff

1. Introduction

Behavioral public policy (BPP) is often treated as a single type, as witnessed, for example, in the popular use of the "nudge" label to encompass all BPP and also in the academic discussion of the pros and cons of BPP *generally*. This has led, first, to an unwarranted polarization in the debate; second, to a neglect of the context sensitivity of these pro and con arguments; and third, to a disregard of multiple stable kinds of policies within the BPP category that could capture these context sensitivities.

Against this *uniformity assumption*, we have argued that the BPP category contains multiple types of policies, distinguished by mechanisms (Grüne-Yanoff and Hertwig 2016; Hertwig and Grüne-Yanoff 2017). Our main argument for this distinction is that there are systematic differences in the context sensitivity of both the effectiveness and the ethical evaluation of these mechanism-based types. Specifically, we claim that there are at least two kinds of behavioral policies, *nudges* and *boosts*, operating through different kinds of mechanisms. We do not claim, however, that these are the only kinds of BPP.

The main purpose of distinguishing the types of BPP by mechanism is to provide a systematic base for the context-sensitive evaluations of their effectiveness and ethical acceptability, thus overcoming the current polarization. The argument therefore is not directed against nudge-type interventions. Instead, it criticizes those who treat BPPs as of one kind, either to universally praise or to universally condemn them. Instead, it is argued that nudge and boost mechanisms have different *moderators*, thus explaining why the respective policy types exhibit different degrees of effectiveness in different contexts and different populations and that they have different potential *side effects*, thus explaining why the respective policy types exhibit different degrees of ethical permissibility in different contexts and different populations. To overcome the polarization and to provide a more powerful tool to analyze which policy type might fare better (either effectively or morally) in which environment is what motivates our categorization proposal.

I start this chapter by sketching the diversity of BPPs (Section 2) and arguing why this diversity matters (Section 3). Section 4 outlines the notion of mechanism used in the analysis. Section 5 develops the distinction between nudges and boosts on the basis of mechanisms and illustrates some of the uses of this categorization. Section 6 concludes.

 DOI: 10.4324/9781315739793-43

2. The Diversity of BPPs

In both economics and psychology, investigations of nonincentivizing and noncoercive determinants of individual human behavior have enjoyed increasing popularity in recent decades. Research of this kind has often been summarized under the label "behavioral" (Heukelom 2014). When these academically driven efforts turned their attention to devising policy recommendations (Jolls et al. 1998; Camerer et al. 2003; Sunstein and Thaler 2003), and policymakers began paying attention (Lunn 2014; Chetty 2015; Geiger 2016), the moniker followed, and "behavioral policy" or "behavioral public policy" became the widely adopted collective term for these recommendations and their implementation. At the same time, due to the popularity of Thaler and Sunstein's (2008) book, "nudges" became a near-synonym for "behavioral policy," and various policy institutions, most prominently the British Behavioural Insights Team (BIT), became known unofficially as "Nudge Units."

Unsurprisingly, perhaps, such attempts at shaping policy have attracted a fair amount of criticism, both from an ethical perspective (for reviews, see Barton and Grüne-Yanoff 2015; Schmidt and Engelen 2020) and from those worried about the effectiveness of the proposed intervention (e.g. House of Lords 2011). At least in the early days, that literature often treated BPP (or the synonymously used nudges) as a single kind, to be either uniformly praised or uniformly condemned. This implicit *uniformity assumption* quite strongly polarized the debate into those rejecting nudges and those endorsing them.

Upon closer inspection, however, one finds a lot of diversity contained within the terms BPP or nudge. Here, I want to emphasize this variety in terms of three dimensions: theory background, various definition attempts, and heterogeneous mechanisms.

First, the theoretical background of behavioral policy recommendations has been quite diverse. For example, although Thaler and Sunstein, in their book *Nudge*, stress their commitment to the heuristics and biases (H&B) tradition initiated by the research of Tversky and Kahneman, many of the policy interventions they describe in fact do not fit well with that tradition. To name but just two examples presented in *Nudge*: one, the famous Amsterdam airport fly in the urinal was developed long before behavioral scientists turned to policy (Kulich 2009), and it is unclear what H&B model would explain its success. The other, the arrangement of stovetop knobs in such a way that they more obviously relate to the burners they control, arose from ergonomics experiments in the 1950s (Chapanis and Lindenbaum 1959). Again, the relation to any H&B model is unclear.

Second, attempts to define nudges broadly as encompassing all behavioral policies have run into various difficulties. Thaler and Sunstein (2008: 6) defined nudges as

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

This definition largely characterizes nudges by what they are *not*: not coercive and not incentivizing. It probably gave rise to the idea that all BPPs are of one kind, contrasted with only coercive and incentivizing interventions. However, some of the interventions discussed in *Nudge* do not fit well even within this very broad definition. For example, the placement of mandatory fuel consumption stickers on the back of cars "for other drivers to see" (Thaler and Sunstein 2008: 194), which produce public shaming effects, although probably an effective intervention, does not square well with incentive neutrality and noncoerciveness. Furthermore, Thaler and Sunstein proceed to qualify this definition by arguing that nudges affect only the behavior of those who are not fully rational, while leaving fully rational agents unaffected (Thaler and Sunstein 2008: 8), implying that nudges operate

by harnessing these irrationalities somehow. This is a considerably narrower characterization than the one cited earlier, leaving room for other, nonnudge interventions in the BPP category. Many authors have struggled to draft more specific definitions of nudges (Bovens 2009; Hausman and Welch 2010; Rebonato 2012; Heilmann 2014; Hansen 2016; Mongin and Cozic 2018), but there is currently no agreement between them.

Third, where authors have investigated the causal pathways through which BPPs operate, this yielded quite diverse results. Some policies operate by removing environmental factors that allegedly influence people to make bad decisions, for example, by banning the sale of supersized soft drink portions or by avoiding the presentation of saving decisions as between "now" and "much later." Other policies operate by encouraging people to rely on their intuitive rules of thumb ("gut instincts") in appropriate circumstances (Gigerenzer 2015). Yet others operate by training people in new heuristics more suitable for the relevant tasks than the ones they are currently using, for example, by representing uncertainty as natural frequencies instead of probabilities when dealing with base-rate-sensitive problems (Hoffrage et al. 2000).

To conclude, BPPs are diverse in a number of dimensions. This is a prima facie reason against treating them as one kind. However, useful kinds often contain a fair amount of diversity as long as it does not defeat the purpose for which the kind is used. In the next section, I will argue that treating BPPs as one kind has such defeating consequences.

3. Consequences of BPP Diversity: Context Dependence

In the previous section, I showed that BPPs are diverse. In this section, I argue that this diversity is problematic because it makes BPPs' effectiveness and ethical evaluation context sensitive. That is, the same intervention is effective and ethically acceptable in one context, but not in another. Such context sensitivity is undesirable, as long as it is not systematically analyzed, because it makes it difficult to anticipate the performance of a policy in a new context.

Consider this example. A municipality might offer consumers a choice of energy providers, setting a slightly more expensive but more sustainable provider as the default. For many consumers, comparing the alternative providers and determining which one is best is an effortful undertaking: they are likely to stick to the default, because they sense that the effort of performing the comparison is higher than the potential gains. In such environments, the municipality's intervention might well be effective in getting these people to choose the green provider. Now imagine instead that, at an earlier point in time, a nongovernmental organization (NGO) had developed a web-based tool that allows a simple but trustworthy comparison of the providers suited for individual consumers' needs. The comparison might become so much less effortful that more consumers will actually perform the comparison and choose accordingly. In such an environment, it is less likely that the municipality's default-setting intervention is effective: people who want green energy at the given price are likely to choose the green provider, and those who do not are likely to choose an alternative, irrespective of how the default was set.

A lot of evidence for such context sensitivities can be found in experimental studies of BPP. Most experiments investigate the *effect size* of an intervention on subjects' behavior, either in a laboratory or in some specific field context. Recorded effect sizes for many behavioral interventions vary widely. Take, for example, information interventions aiming to reduce household energy consumption. The most recent meta-analysis in this domain, covering 156 studies, found a weighted average treatment effect of −7.40%, that is, on average, information interventions produced more than 7% in potential savings (Delmas et al. 2013). However, the range of individual effect sizes varied from −55% to +18.5%! Possible explanations for this wide range of results include study quality (high-quality studies, according to the meta-analysis, found lower effect sizes than low-quality studies) and differences in what information the intervention provided (e.g. consequences of high energy

consumption, suggestions for how to lower consumption, the consumer's own past consumption, others' consumption), as well as contextual variations: who the subjects were and in what context they received the information. For example, in a highly politicized context, information about the consequences of behavior is often discounted along partisan lines (Tannenbaum et al. 2017); procedural information about energy savings will have little impact if a subject does not have access to the necessary technology; and information about others' consumption tends to have a higher impact if they are part of the subject's peer network (Gächter et al. 2013). Context sensitivity means that the effect size of the intervention depends on the presence or absence of such contextual factors as politicization, access to technology, and peer networks.

Information policies are not equally sensitive to all of these factors, though. Information about consequences is not likely to be sensitive to technological access, nor is procedural information to peer network. This is because the information provided by these interventions affects behavior through different *causal pathways*: one affects evaluations, another instrumental beliefs, and a third social norm conformity. All of these policies employ the same *intervention lever*: they provide relevant information to subjects. Nevertheless, these policies need to be further differentiated by the mechanism through which they operate. Only by distinguishing them by mechanism does it become clear why different information policies are sensitive to some contextual factors but not to others. To make unambiguous claims about a policy's effectiveness in certain contexts, one needs to determine whether the contextual factors impact the intervention's effect size. This requires the identification of the mechanism through which it operates (see also Clarke, Chapter 21).

Knowledge about a policy's mechanism is also important for assessing its ethical acceptability (Smith et al. 2013; Grüne-Yanoff 2016). For example, it might be important to know, for such an assessment, whether an intervention like the preceding green default setting is transparent to the subject. This requires insight into how it operates on subjects, for example, subconsciously or by signaling some relevant information. But this again requires knowledge about the intervention mechanism, which is not available from the mere categorization by intervention lever.

Thus, BPPs' effectiveness and ethical acceptability often depend on the context in which they are implemented. Simply accepting such context sensitivity is not a viable option. Policymakers need to make ex ante judgments about the effectiveness and ethical acceptability of intervention alternatives in target environments and for target audiences. If the specific policy has already been tested in that environment, judgment can be passed with some confidence. But most policies have not been tested in their target environment, and the performance of a serious test would be prohibitively expensive or cause significant delay (also note the difficulty of selecting what to test). The policymaker thus faces the problem of *extrapolation* (Steel 2008; Cartwright 2012): an intervention I is found to be effective and ethically acceptable in context C, but it is unclear whether it will be in context D. If one cannot examine I in D directly, then one must link I to D in other ways, which requires some form of generalization and categorization (e.g. "Interventions of type $\boldsymbol{T}$ tend to perform like this in D. I is a $\boldsymbol{T}$. Therefore, expect I to perform like this in D"). For this reason, ex ante judgments require categorization. A naive plea for policy assessment on a "case-by-case" basis (e.g. Sims and Müller 2019) founders on this observation.

But not just any categorization works. Current practices seem to sometimes use intervention levers as relevant subcategories for different BPPs.[1] This is an entirely plausible strategy in the early stages of a research field. However, for the purposes of evaluating interventions, such a categorization is not sufficiently stable, for at least two reasons. First, as I already indicated, interventions with the same lever might operate through different mechanisms. Second, one often cannot even properly determine the intervention lever without knowing through which mechanism the policy operates. The answers to questions like, "Does the intervention provide information or set a frame?" or "Does it offer incentives or change the choice architecture?" require reference to at least some features of the causal pathways through which the intervention operates.

Categorization by levers does not help with context sensitivity. Because the same lever might trigger different mechanisms, each of these mechanisms might be affected differently by contextual variables, both in terms of preventing or amplifying the intervention's effect and also in terms of preventing or producing various side effects. Thus, attempts to categorize policies characterized merely by their intervention levers as more or less ethically acceptable (Oliver 2013; Baldwin 2014) are hopeless, as it is the population and the environment that at least partially determine ethical evaluation, which thus undermines such simple classification attempts.

Instead, we have suggested the categorization of BPPs according to the mechanisms through which they operate and then, on the basis of this categorization, infer the transferability of an intervention's effectiveness and ethical acceptability from one context to another (Hertwig and Grüne-Yanoff 2017; Grüne-Yanoff et al. 2018). The examples we discussed in this section make this an intuitive solution: the context sensitivities exemplified here depend on the causal pathways through which the default-setting interventions operate. In order to give a more general analysis that justifies our proposal, I need to discuss the notion of mechanism in more detail.

4. Systematizing Diversity: A Mechanistic Account

The current philosophy of science characterizes *mechanisms* broadly as systems of causally interacting parts and processes, which under certain conditions predictably produce one or more effects (e.g. Craver and Tabery 2019; Glennan 2017). For BPPs, the relevant mechanisms link intervention levers to agents' behavior. The link between these components is mediated by the agents' decisions and the environment in which these decisions are taken.

Many authors understand talk of mechanisms as talk about elements of the real world. A grandfather clock's mechanism consists of the actual pendulum, spring, and gears. But models that represent such components either fully or partially (e.g. different number of teeth, but same ratio in the gear train) are not considered mechanisms in this *ontic* sense. In contrast, I consider mechanisms to comprise abstracting models, for three reasons. First, in the behavioral sciences, there is little agreement about the correct level of description. Decision mechanisms can often be described on both the social and the individual-mental levels, and sometimes the neurological level. While these levels typically supervene each other, multiple realizabilities complicate attempts at reduction and thus leave open the question of whether any level is ontologically prior to the others. Second, even if one fixes the level of description, there is uncertainty how fine grained the individuation of mediators should be. For example, should one describe the cognitive-cost-based default intervention as operating though one mediator ("cost-benefit-assessing module"), or should this be unpacked into a sequence of observations, belief formations, comparisons, and evaluations? Because the behavioral sciences lack an atomistic framework for their ontology (contrast this, for example, with the molecular level that biochemists can refer to), any "more fine grained is better" strategy runs into the issue of dividing ad infinitum without any clear benefit. Finally, besides these ontological worries, there also are epistemic considerations that speak against relying on an ontic conception of mechanism for the purpose of categorizing BPP. It is difficult to obtain evidence for behavioral mechanisms, and the more fine grained the description of the mechanism, the more pressing the problem of underdetermination by the evidence. Therefore, also for epistemic reasons, it is often advisable to rely on abstract mechanistic models instead of an ontic conception of mechanisms.

But what is the right level of abstraction, then? This depends on the model user's epistemic and pragmatic interests. For behavioral policymaking, mechanisms are chiefly of interest because they hold important information about what factors further or inhibit the policy's effectiveness and what side effects a policy might have in certain environments (Grüne-Yanoff 2016; Marchionni and Reijula 2019).

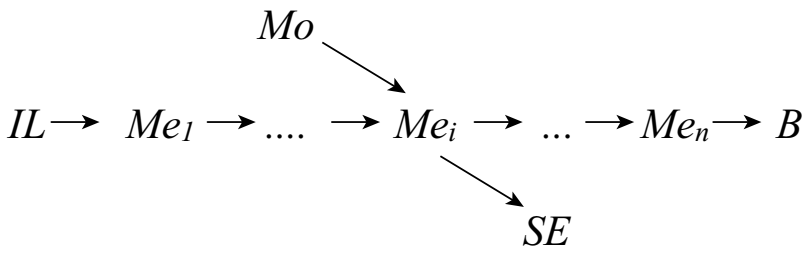

Figure 35.1 Mechanisms in behavioral policy making

Figure 35.1 describes the schematic form of such mechanisms, using the following terminology. A *behavioral policy* consists of an *intervention lever* (*IL*) that the policymaker cranks with the intention of effecting some change in individual *behavior* (*B*). The causal chain from intervention level to behavior can be represented as consisting of a sequence of *mediators* (Me_i). A mediator can either pass on a causal signal or block it. This depends on *modulators* (*Mo*): factors that affect mediators. Besides passing on causal signals to their successor in the causal chain, mediators also might have *side effects* (*SE*).

To illustrate, the provision of procedural information is an *IL* that might change an instrumental belief about thermostat settings, leading to the intention to set the thermostat 2 degrees lower when absent (Me_i). But if there is no thermostat in the apartment (*Mo*), this intention cannot be implemented. The modulator *Mo* thus prevents the effect of the intervention on consumption behavior *B*.

To give another illustration, the default setting of a green energy provider is an *IL* that might make the subject feel that a comparison is too costly given the potential gains (Me_i) and thus lead her to stick to the default option (*B*). The provision of a simple and trustworthy comparison tool might reduce costs to such an extent that the effect of *IL* on *B* is blocked. If it is not blocked (*Mo* absent), the elicitation of such a feeling might contribute to a general sense that bureaucratic communications are not worth serious consideration (*SE*), and such a side effect might be important when assessing the effectiveness and ethical dimensions of such interventions.

This mechanistic account helps to make precise the analysis of context sensitivity from the previous section. To evaluate the effectiveness of an *IL*, in a given context *C*, the mechanism through which *IL* affects *B* determines which modulators Mo_i must be either present or absent. By checking whether *C* contains these Mo_i we will be able to draw justified conclusions about the effectiveness of *IL* in *C*. To assess the ethical acceptability of an *IL*, knowledge of the mechanism allows us to check whether its operation, through specific Me_i or having particular *SE*, is ethically problematic.

Admittedly, policymakers often do not know the exact mechanisms through which their BPP options might operate. But if full knowledge is not attainable, a second-best option of knowing through which mechanism *kind* a BPP operates still serves the same purpose of assessing effectiveness and ethical acceptability. But how could one meaningfully distinguish between different kinds of BPP mechanisms? To this question I now turn.

I will start with an abstract and highly simplified mechanism scheme of decision-making, depicted in Figure 35.2. Individual decision-makers distinguish a number of *alternatives*, identify their relevant *properties* (e.g. their possible consequences and the uncertainty with which they come about), and choose one of the alternatives according to some *selection rule*. This rule might involve the *evaluation* of consequences and their uncertainty, but it might also be a rule as simple as "choose the alternative highest up on the list."

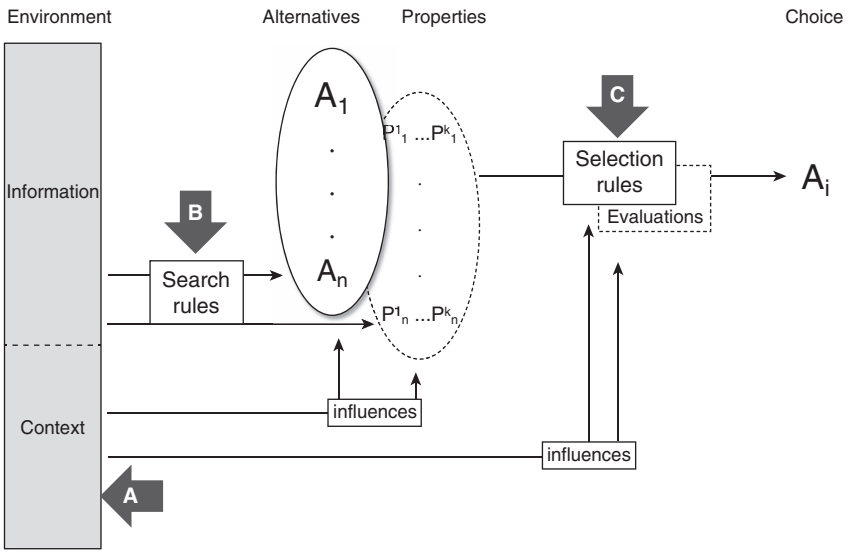

Figure 35.2 A simplified mechanism scheme of decision making

But how do individuals arrive at a list of alternatives and their relevant properties? They search the environment according to some *search rules* that tell them what information to focus on, how to mark distinctions, and when to stop searching. This also means that individuals consider only some features of the environment as relevant, while ignoring others. The search rule therefore divides the decision environment into relevant *information* and irrelevant *context*. What behavioral research has shown, however, is that context considered irrelevant by individuals might nevertheless have causal impact on their decision. Some such contextual factors as, for example, "anchors" or "frames," purportedly influence the representation of alternatives and their properties; other factors like "default effects" or "reference points purportedly influence the evaluation and selection of alternatives, although the individual considers them irrelevant.

This scheme describes ways in which individuals make decisions (albeit, as mentioned, in a simplified way). Now we can identify various points at which a BPP intervention might attack.

The large dark gray arrows in Figure 35.2 indicate different possible interventions on these decision mechanisms.[2] Intervention A intervenes in the contextual factors, removing, rearranging, or adding some, with the intention of exerting influence through them on the individuation and characterization of alternatives and on the search rule and the underlying evaluation. For example, the "Save More Tomorrow" intervention offers deliberators the choice between more consumption *in a year's time* and higher pension payouts *later*, thus avoiding the "present bias" impact of a choice of more consumption *now* on the evaluation of the alternatives (Thaler and Benartzi 2004).

Interventions B and C, in contrast, intervene in the search and selection rules directly, by teaching people new skills or training existing ones. For example, Finkel et al.'s (2013) intervention to reduce marital strife trains people a new selection rule ("assume the perspective of a third-party spectator"). Drexler et al.'s (2014) physical accounting intervention, in contrast, teaches people with little formal education a better search rule for their business purposes ("physically separate private and business receipts"). The first difference between these policy mechanisms thus consists of the location of the *entry point* for the intervention lever.

The second difference consists of the different mediators that connect the intervention lever with the behavior. "Save More Tomorrow," for example, operates through the mechanism that underlies

the present bias, while Drexler et al.'s physical accounting intervention operates through changing the search rule. Sometimes the same intervention lever can operate though different mechanisms. Default-setting policy interventions, for example, have been speculated to operate though either cognition-cost avoidance, status-quo bias, or receiving recommendation signals (Grüne-Yanoff 2016; Jachimowicz et al. 2019). But in those cases, in order to determine *where* the intervention lever comes in, one must refer to the mechanism: a loss-aversion-driven, default-setting policy, for example, comes in through an intervention on context factors, while a recommendation-driven, default-setting policy operates through enlarging the searchable information set and, thus, comes in through an intervention on relevant information.

Third, policies also differ in the moderators that might inhibit their operation. Default setting driven by cognition costs, as I argued in Section 2, is sensitive to changes in cognition costs, while default setting driven by status-quo bias is not. "Save More Tomorrow" operates through the mechanism that shapes the intertemporal discounting curve hyperbolically and thus produces present bias. If that curve were to change under the intervention, then the policy would not be effective. Finkel et al.'s marital strife intervention would be blocked if people did not want to end an altercation, even though the policy had now taught them how to do it. And Drexler et al.'s physical accounting intervention likely would not be effective if the lacking business discipline was not caused by the inability to extract relevant information but by, for example, rampant corruption. All of these factors are examples of modulators that reduce the effectiveness of those behavioral policies, whose mediators they block. Policies that operate through other kinds of mediators, in contrast, will not be affected by those factors.

Fourth, policies that differ in mediators also might differ in their side effects. For example, a default-setting intervention operating through a cognitive-cost mechanism might leave the people thus affected with the general impression that bureaucracy communications are hard to comprehend and not worth the effort (after all, the intervention must elicit this impression for its specific communication to be effective). In contrast, default-setting interventions operating through recommendation or loss-aversion mechanisms are less likely to have such a side effect. Similarly, a normative feedback intervention operating through a social pressure mechanism might induce people thus affected to hide their behavior from public view (and thus from potential sanctions), while a normative feedback intervention operating through a reference point mechanism is unlikely to cause such side effects.

BPP mechanisms thus can be systematically distinguished with respect to at least these four criteria. With this, I now turn to the proposed distinction between boosts and nudges.

5. Boost vs. Nudge

Nudges and boosts have been characterized in multiple dimensions, some of which explicitly distinguish the causal pathways through which these two types of interventions operate (Hertwig and Grüne-Yanoff 2017: 974). Nudges "harness cognitive and motivational deficiencies in tandem with changes in the external choice architecture"(Ibid.). Boosts, in contrast, "foster competences through changes in skills, knowledge, decision tools, or external environment" (Ibid.). The distinction thus rests on two criteria. First, *where* the intervention lever is applied: nudges intervene in the choice architecture, while boosts intervene in skills, knowledge, decision tools, or external environment. Nudge intervention entry points thus largely correspond to intervention A in Figure 35.2, while boosts largely correspond to intervention entry points B and C, but they sometimes also attack through A.

The second criterion is *how* the intervention affects behavior: nudges harness cognitive and motivational deficiencies, while boosts foster competences. Nudge interventions typically operate by harnessing factors that the decision-maker herself regards as irrelevant but that have been shown

to have an influence nevertheless, This corresponds to the pathways in the lower part of Figure 35.2. Boost interventions, in contrast, typically operate by affecting competences that the decision-maker considers relevant, for example, by providing new competences or by better matching existing competences with the task at hand, This corresponds to the pathways in the upper part of Figure 35.2. This also helps to distinguish those boosts that intervene on the external environment and thus have A as their entry point. In contrast to nudges, which attack at A in order to harness factors that are effective but disregarded by the decision-maker, boosts intervening at A seek to turn a disregarded factor into one that the decision-maker takes into account and thus increases her competence. For example, an intervention that translates statistical information from a relative probability format into a natural frequency format helps decision-makers become aware of the framing effects these formats have on their decisions and, through that, improve their competences. A nudge, in contrast, would choose to present the information in that format which is expected to yield the desired framing effect, without necessarily teaching the decision-maker any competence.

The mechanisms supporting this categorization are highly abstract models. Why this level of abstraction? The preceding epistemic considerations give at least a partial answer: to distinguish mechanisms (and hence BPP categories) only to the degree to which such distinction is supported by evidence, either directly from experiments or other empirical studies or indirectly through empirically supported background theory. But this is only a partial answer. The BPP categorization is determined not only by available evidence but also by the purposes that such a categorization may have. The recent literature gives clear indications of what pragmatic considerations make researchers and policymakers turn to mechanisms (Hertwig 2017; Grüne-Yanoff et al. 2018; Strassheim 2019; Löfgren and Nordblom 2020): to provide resources for ex ante evaluations of interventions' effectiveness and ethical acceptability. Reference to mechanisms facilitates such ex ante evaluations because it helps to identify which factors in the intended context of implementation make a difference to the effectiveness and the side effects of the intervention. What matters, therefore, is that the mechanisms by which BPPs are categorized are sufficiently fine grained to flag such difference-making contextual factors.

Different BPP interventions are sensitive to some contextual factors but not to others. Categorization of BPPs into nudges and boosts collects BPPs sensitive to some factors in one category and all those not sensitive to those factors in another. This is because the mechanisms on which these categories are built either have that factor as a modulator or they do not. Such a categorization is helpful for the policymaker in at least two ways. First, she only needs to worry about the contextual factors to which her policy category of interest is sensitive. Second, she now knows to which factors her policy category of interest is sensitive, so she can go and check whether these factors are active in the target environment. Let us consider a few such factors now (Table 35.1).[3]

Nudges operate by harnessing cognitive and motivational heuristics. The cognitive-cost version of default setting, for example, relies on the subject's feeling that the choice alternatives are not worth looking at, thus making her stick to the default. This reliance makes the intervention sensitive to any factor that destabilizes such a feeling, be it a simple app that allows a cheap and reliable comparison or a prod from a trusted friend. Such modulators undermine the *heuristic stability* of nudge

Table 35.1 Context conditions of Nudges and Boosts

Context conditions	*Nudges*	*Boosts*
Heuristic stability	☑	–
Agent motivation	–	☑
Teachability of heuristics	–	☑
Homogeneity of heuristic repertoire	☑	–

mechanisms, thus rendering them less effective. Boosts, in contrast, because they do not harness deficiencies in this way, are not sensitive to this kind of modulator. Thus, policymakers intending to implement a nudge in a specific context are well advised to check whether it contains such modulators; policymakers intending to implement a boost do not need to worry.

Boosts operate by fostering competences, which might, for example, consist of search or decision rules more suitable for a given task. But such interventions will only have an effect on behavior if the subject, when addressing the task, is actually motivated to choose these newly acquired or newly identified rules. Factors that reduce or eliminate *agent motivation* thus are modulators of boost mechanisms. A boost, even if it succeeds in teaching a new competence, will not be effective if the agent is not motivated to use what the boost taught her. Nudges, in contrast, because they do not operate through the motivation of agents, are not sensitive to this kind of modulator. Thus, policymakers intending to implement a boost in a specific context are well advised to check whether it contains such modulators; policymakers intending to implement a nudge do not need to worry.

The way boosts foster competences is by teaching skills, knowledge, decision tools, etc. But what if these cannot be taught? Some modes of human cognition or perception are not very malleable. One cannot teach humans to "see" that the two lines in the Müller-Lyer illusion are of equal length (Fodor 1983); even if people have convinced themselves otherwise, they will still see them as unequal in length. Furthermore, the policymaker might fail to teach people even about malleable features, if contextual factors causing distraction, inattention, or inability are present. Such factors undermine the *teachability of heuristics* and, thus, constitute modulators of any boost mechanism that seeks to teach them. Nudges, in contrast, because they do not operate through teaching agents, are not sensitive to this kind of modulator. Thus, policymakers intending to implement a boost in a specific context are well advised to check whether it contains such modulators; policymakers intending to implement a nudge do not need to worry.

Nudges and boosts are interventions that target individuals. But realistically, both are most often applied to large populations where, for example, individual therapeutic approaches are not feasible. Thus, nudges intervene in the choice architecture of all agents in the population, and boosts intervene in the skills, knowledge, decision tools, etc. of all of them. This poses more of a problem for nudges than for boosts. Imagine a nudger "reframing" the description of an unhealthy product, expecting that the new frame would signal its riskiness and thus deter those susceptible to this signal. The nudger thus implicitly assumes that individuals either pick up the signal and react in the desired way (i.e. consume less of the product) or are not susceptible to the signal. Yet, what if some in the population are susceptible to the signal but react to it by consuming more? In such a heterogeneous population, it becomes unclear what effect the nudge has; it might have no effect, or the effect might be the opposite of what was expected. Factors that undermine the *homogeneity of the heuristic repertoire* are modulators of nudge mechanisms. Boosts, in contrast, do not require this homogeneity of the population. To the extent that agents already know what the boost intervention teaches them, they can safely ignore it. To the extent that it teaches something new, it is the agents' choice to make use of it, and thus little homogeneity (beyond basic learning abilities) is required. Thus, boosts are not sensitive to this kind of factor. Policymakers intending to implement a nudge in a specific context are well advised to check whether it contains such modulators; policymakers intending to implement a boost do not need to worry.

My discussion so far has focused on how mechanism-based categorization can help to deal systematically with the context sensitivity of BPP interventions' effectiveness. Now, I briefly discuss some ways in which mechanism-based categorization can also help to deal systematically with the context sensitivity of BPP interventions' ethical acceptability.

I will start with *transparency*, which is widely regarded as an ethically relevant feature of behavioral policy (Bovens 2009). Policy is transparent only if people affected by the intervention can easily learn about the factors influencing them. Boosts guarantee such transparency. A boost seeks to

impart a competence by teaching skills, knowledge, decision tools, etc. But such teaching requires the cooperation of the subject to be influenced: she must pay attention when taught, she must grasp the content she is taught, she must be aware of what these skills and tools can be used for, and she must elect to use them at some point for the boost to have had any effect on her behavior. This might sound more involved than it is. To teach the simple third-party perspective of the marital strife intervention (Finkel et al. 2013), for example, will not require more than a few minutes of attention, comprehension, and awareness, nor does it require huge cognitive effort from the subjects. But it is hard to imagine how an individual could cooperate in these ways without learning about the factors influencing them. Transparency is, in this sense, built into the boost such that people affected by it are aware of it its inevitable side effect. That is not the case with nudges. Although nontransparency might not be necessary for a nudge to be effective (Loewenstein et al. 2015), transparency certainly is not necessary for its effectiveness and is not its inevitable side effect. A change in the choice architecture is often effective even if the influenced agents are not aware of it. Thus, nudges require more ethical scrutiny than boosts because they can circumvent transparency, while boosts cannot.

A second ethical issue is to what extent BPPs affect the *autonomy* of decision-making (Hausman and Welch 2010). The philosophical debate about what constitutes autonomous decisions is, of course, enormous, so I will simply pick one prominent account, coherentism, to illustrate how mechanism-based categorization facilitates a systematic discussion. According to coherentist views, an individual has control over her own action, if and only if she is motivated to act in this way because this motivation coheres with some mental state that represents her point of view on the action (Buss and Westlund 2018). There is little agreement on what these relevant mental states are. However, even without specifying that, a consideration of BPP mechanisms can help clarify the debate. The agent's motivation is what causes her to act. She can endorse these causes (in that case the mental states representing her point of view cohere with them) or she can repudiate them. To have reasonable grounds for repudiation, the agent must experience some disconnect between her motives and her point of view. Someone or something might force her to act in this way, for example, or she must consider some of her own motivations are "not really her own." It would not be reasonable for her to repudiate motivations that she formed without external pressure and absent any internal conflict (see also Lecouteux, Chapter 4).

Boosts operate through imparting skills, knowledge, decision tools, etc. These interventions are not effective without the individual's cooperation. She must elect to make use of the thus imparted competences; the fact that she learned a new skill, for example, has no effect on her behavior unless she is capable and motivated to apply it. Unless the agent repudiates this motivation, the application of the boost constitutes an autonomous decision. Such repudiation is not likely, for at least two reasons. First, genuine boosts do not impose pressure on agents to use the competences they impart, so the individual has little reason to repudiate her motivation for that reason. Second, motivations for applying a boosted competence arise from reflections of what might be best for the agent in this situation, and typically they are neither impulsive or subtly seductive. Therefore, the individual has little reason to repudiate her motivation for those reasons either. Consequently, *due to the boost mechanisms*, it is very unlikely that the application of boosted competences would constitute heteronomous decisions (Grüne-Yanoff 2018).

This stands in marked contrast to nudge mechanisms. Nudges operate through causal pathways that might circumvent agential reflection or that might overrule existing motivation. They often proceed through mechanisms parallel to and independent of the deliberative processes that facilitate reflection about one's motivations. For these reasons, it is likely (but of course not necessary) that that nudges produce motivation that the agent repudiates as not her own and that therefore constitute heteronomous decisions. Nudges thus require more ethical scrutiny than boosts, because their mechanisms are more likely to produce heteronomous decisions (according to coherentist views) than boosts.

6. Conclusion

Although BPP interventions are often treated as if they form one category, members of this category are actually rather diverse. A consequence of this diversity is not only that BPPs differ in their effectiveness and ethical evaluations but that their effectiveness and ethical evaluations are context dependent. For a systematic treatment of their context-dependent performance, BPPs should be categorized according to the mechanisms through which they operate. In particular, BPPs should be distinguished into two categories, nudges and boosts. Relevant conclusions about effectiveness and ethical acceptability of a BPP in novel contexts can be derived from this mechanistic distinction between nudges and boosts, thus offering the policymaker a strategy to deal with the problem of extrapolation.

Acknowledgments

I thank Julian Reiss and Ralph Hertwig for very helpful comments on an earlier draft of this chapter.

Related Chapters

Clarke, Chapter 21 "Causal Contributions in Economics"
Lecouteux, Chapter 4 "Behavioral Welfare Economics and Consumer Sovereignty"

Notes

1 That the literature categorizes BPPs according to such levers is not surprising. A lot of effort in behavioral research has gone into establishing standardized experimental designs, through which stable phenomena or stable intervention effects can be established (Guala 2005). Consequently, the different intervention proposals have often been categorized according to these standardized experimental manipulations (e.g. "default-setting," "framing," or "feedback" designs).
2 There are other possible interventions not illustrated here, for example, information campaigns enlarge searchable information and coercion eliminates some alternatives, while incentives change some of the alternatives' properties.
3 For a more comprehensive treatment of such factors, see Hertwig and Grüne-Yanoff (2017), Hertwig (2017), Grüne-Yanoff et al. (2018), Grüne-Yanoff (2018).

Bibliography

Baldwin, R. (2014) "From Regulation to Behaviour Change: Giving Nudge the Third Degree," *The Modern Law Review* 77(6): 831–857.
Barton, A., and Grüne-Yanoff, T. (2015) "From Libertarian Paternalism to Nudging – and Beyond," *Review of Philosophy and Psychology 6*(3): 341–359.
Bovens, L. (2009) "The Ethics of Nudge," in T. Grüne-Yanoff and S.O. Hansson (eds.) *Preference Change* (pp. 207–219), Dordrecht: Springer.
Buss, S., and Westlund, A. (2018) "Personal Autonomy," in Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/.
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.
Cartwright, N. D. (2012) "Presidential Address: Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps," *Philosophy of Science 79*(5): 973–989.
Chapanis, A., and Lindenbaum, L.E. (1959) "A Reaction Time Study of Four Control-Display Linkages," *Human Factors 1*(4): 1–7.
Chetty, R. (2015) "Behavioral Economics and Public Policy: A Pragmatic Perspective," *American Economic Review 105*(5): 1–33.

Craver, C., and Tabery, J. (2019) "Mechanisms in Science," in Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), https://plato.stanford.edu/archives/sum2019/entries/science-mechanisms/.

Delmas, M. A., Fischlein, M., and Asensio, O. I. (2013) "Information Strategies and Energy Conservation Behavior: A Meta-Analysis of Experimental Studies from 1975 to 2012," *Energy Policy 61*: 729–739.

Drexler, A., Fischer, G., and Schoar, A. (2014) "Keeping It Simple: Financial Literacy and Rules of Thumb," *American Economic Journal: Applied Economics 6*(2): 1–31.

Finkel, E. J., Slotter, E. B., Luchies, L. B., Walton, G. M., and Gross, J. J. (2013) "A Brief Intervention to Promote Conflict Reappraisal Preserves Marital Quality Over Time," *Psychological Science 24*: 1595–1601.

Fodor, J. A. (1983) *The Modularity of Mind*, Cambridge, MA: MIT Press.

Gächter, S., Nosenzo, D., and Sefton, M. (2013) "Peer Effects in Pro-Social Behavior: Social Norms or Social Preferences?" *Journal of the European Economic Association 11*(3): 548–573.

Geiger, N. (2016) "Behavioural Economics and Economic Policy: A Comparative Study of Recent Trends," *Œconomia. History, Methodology, Philosophy* (6–1): 81–113.

Gigerenzer, G. (2015) *Bauchentscheidungen: Die Intelligenz des Unbewussten und die Macht der Intuition*, München: C. Bertelsmann Verlag.

Glennan, S. (2017) *The New Mechanical Philosophy*, Oxford: Oxford University Press.

Grüne-Yanoff, T. (2016) "Why Behavioural Policy Needs Mechanistic Evidence," *Economics and Philosophy* 32(3): 463–483.

Grüne-Yanoff, T. (2018) "Boosts vs. Nudges from a Welfarist Perspective," *Revue d'Économie politique* 128(2): 209–224.

Grüne-Yanoff, T., and 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. (2018) "Toward a Framework for Selecting Behavioural Policies: How to Choose Between Boosts and Nudges," *Economics and Philosophy* 34(2): 243–266.

Guala, F. (2005) *The Methodology of Experimental Economics*, Cambridge: Cambridge University Press.

Hansen, P. G. (2016) "The Definition of Nudge and Libertarian Paternalism: Does the Hand Fit the Glove?" *European Journal of Risk Regulation* 7(1): 155–174.

Hausman, D. M., and Welch, B. (2010) "Debate: To Nudge or Not to Nudge," *Journal of Political Philosophy 18*(1): 123–136.

Heilmann, C. (2014) "Success Conditions for Nudges: A Methodological Critique of Libertarian Paternalism," *European Journal for Philosophy of Science 4*(1): 75–94.

Hertwig, R. (2017) "When to Consider Boosting: Some Rules for Policy-Makers," *Behavioural Public Policy 1*(2): 143–161.

Hertwig, R., and Grüne-Yanoff, T. (2017) "Nudging and Boosting: Steering or Empowering Good Decisions," *Perspectives on Psychological Science 12*(6): 973–986.

Heukelom, F. (2014) *Behavioral Economics: A History*, Cambridge: Cambridge University Press.

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000) "Communicating Statistical Information," *Science 290*: 2261–2262. doi:10.1126/science.290.5500.2261

House of Lords, Science and Technology Select Committee (2011) *Behaviour Change* (Second report), London: House of Lords.

Jachimowicz, J. M., Duncan, S., Weber, E. U., and Johnson, E. J. (2019) "When and Why Defaults Influence Decisions: A Meta-Analysis of Default Effects," *Behavioural Public Policy 3*(2): 159–186.

Jolls, C., Sunstein, C.R., and Thaler, R. H. (1998) "A Behavioral Approach to Law and Economics," *Stanford Law Review* 50(5): 1471–1550.

Krulwich, R. (2009) "There's a Fly in My Urinal," *NPR*, December 19. www.npr.org/templates/story/story.php?storyId=121310977. Last accessed June 12, 2020.

Loewenstein, G., Bryce, C., Hagmann, D., and Rajpal, S. (2015) "Warning: You Are About to Be Nudged," *Behavioral Science & Policy 1*(1): 35–42.

Löfgren, Å., and Nordblom, K. (2020) "A Theoretical Framework of Decision Making Explaining the Mechanisms of Nudging," *Journal of Economic Behavior & Organization 174*: 1–12.

Lunn, P. (2014) *Regulatory Policy and Behavioural Economics*, Paris: OECD. doi:10.1787/9789264207851-en

Marchionni, C., and Reijula, S. (2019) "What Is Mechanistic Evidence, and Why Do We Need It for Evidence-Based Policy?" *Studies in History and Philosophy of Science Part A 73*: 54–63.

Mongin, P., and Cozic, M. (2018) "Rethinking Nudge: Not One but Three Concepts," *Behavioural Public Policy 2*(1): 107–124.

Oliver, A. (2013) "From Nudging to Budging: Using Behavioural Economics to Inform Public Sector Policy," *Journal of Social Policy 42*(4): 685–700.

Rebonato, R. (2012) *Taking Liberties: A Critical Examination of Libertarian Paternalism*, London: Palgrave Macmillan.

Schmidt, A. T., and Engelen, B. (2020) "The Ethics of Nudging: An Overview," *Philosophy Compass 15*(4): e12658.

Sims, A., and Müller, T. M. (2019) "Nudge Versus Boost: A Distinction Without a Normative Difference," *Economics & Philosophy 35*(2): 195–222.

Smith, N. C., Goldstein, D. G. and Johnson, E. J. (2013) "Choice Without Awareness: Ethical and Policy Implications of Defaults," *Journal of Public Policy & Marketing 32*(2): 159–172.

Steel, D. (2008) *Across the Boundaries: Extrapolation in Biology and Social Science*, Oxford: Oxford University Press.

Strassheim, H. (2019) "Behavioural Mechanisms and Public Policy Design: Preventing Failures in Behavioural Public Policy," *Public Policy and Administration*. doi:10.1177/0952076719827062.

Sunstein, C. R., and Thaler, R. H. (2003) "Libertarian Paternalism Is not an Oxymoron," *The University of Chicago Law Review*: 1159–1202.

Tannenbaum, D., Fox, C. R., and Rogers, T. (2017) "On the Misplaced Politics of Behavioural Policy Interventions," *Nature Human Behaviour 1*(7): 1–7.

Thaler, R. H., and Benartzi, S. (2004) "Save More Tomorrow™: Using Behavioral Economics to Increase Employee Saving," *Journal of Political Economy 112*(S1): S164–S187.

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