Economics and Philosophy, page 1 of 21 © Cambridge University Press
doi:10.1017/S0266267115000425
journals.cambridge.org/eap

WHY BEHAVIOURAL POLICY NEEDS MECHANISTIC EVIDENCE

TILL GRÜNE-YANOFF*

Abstract: Proponents of behavioural policies seek to justify them as 'evidence-based'. Yet they typically fail to show through which mechanisms these policies operate. This paper shows – at the hand of examples from economics and psychology – that without sufficient mechanistic evidence, one often cannot determine whether a given policy in its target environment will be effective, robust, persistent or welfare-improving. Because these properties are important for justification, policies that lack sufficient support from mechanistic evidence should not be called 'evidence-based'.

Keywords: Policy, Mechanisms, Nudge, Evidence, Welfare

1. INTRODUCTION

Behavioural policies are interventions in the choice environment that aim at affecting behaviour without incentive change or coercion. Instances of these have become widely known as 'nudges'. These policy proposals are typically based on evidence coming from the behavioural sciences, and they are claimed to be subject to the standards of evidence-based policy. For example, David Halpern, the director of the Behavioural Insights Team in the UK, recently argued just that, in an article entitled 'Nudge unit: our quiet revolution is putting evidence at heart of government' (Halpern 2014).

However, the evidence cited for the justification of such policies tends to be of a particular kind: it shows that the policy intervention, in a particular environment, makes a difference to a behavioural variable of interest. Such evidence is either produced in controlled experiments, where the average effect conditional on the intervention is compared

* Department of Philosophy and the History of Technology, Royal Institute of Technology (KTH), Brinellvägen 32, 10044 Stockholm, Sweden. Email: gryne@kth.se. URL: http://people.kth.se/~gryne/

with the average effect conditional on non-intervention; or alternatively through observational studies, where the average effect is estimated through statistical analysis (Cartwright 2009a; Clarke *et al.* 2014). What is typically missing is any evidence about the underlying mechanisms through which these policies affect behaviour.

In this paper, I argue that this deficit is a serious defect. Without information about the underlying mechanisms, we often are unable to determine the effectiveness, robustness, persistence and welfare properties of these policies in their intended target environments. Consequently, policies without sufficient mechanistic evidence should not be considered 'evidence-based'.

My argument proceeds as follows. Section 2 sketches two examples of behavioural policies. Section 3 discusses different mechanistic models that have been proposed as possible explanations of these policies. Section 4 argues that the effectiveness, robustness, persistence and welfare properties of these policies depend on which mechanistic models we assume to be the correct ones. Section 5 concludes.

2. BEHAVIOURAL POLICIES

Amongst the menu of options that agents choose from, the *default* is the option that obtains if the chooser does nothing. Empirical studies show that setting a default significantly affects how people choose. This is called the *default effect*. Setting or changing defaults has been proposed as an effective way of influencing behaviour, for example organ donation choice (Johnson and Goldstein 2003) or product option selection (Park *et al.* 2000). Because it purportedly affects behaviour without substantial incentive change or coercion, default setting is widely considered a form of nudge policy (Thaler and Sunstein 2008: 83–7).[1]

To illustrate, a recent study shows that by changing the default contribution to a retirement plan from 3% to 6%, the number of new employees who would make the higher contribution could be doubled. Under the old retirement plan, employees who did not make an explicit choice would be enrolled in a 3% contribution scheme; under the new policy, they would be enrolled in a 6% contribution plan. As a result of the new policy, 49% rather than 24% ended up with higher contribution rates (Beshears *et al.* 2009). More generally, proponents have argued that by setting the default to that option which is considered optimal, more people can be induced to choose the optimal option.

[1] However, because default settings are sometimes seen as recommendations, and hence as information provisions, some have argued that default settings are not necessarily nudges (for a discussion, see Grüne-Yanoff and Hertwig 2015).

Another behavioural policy is the *Save More Tomorrow*$^{\text{TM}}$ intervention proposed by Thaler and Benartzi (2004, 2013). The goal of this policy is to increase savings contribution rates amongst continuing employees. For this purpose, a policy intervention with three components is implemented:

> First, employees are invited to commit now to increase their saving rate later, perhaps next January or a few months in the future. Self-control is easier to accept if delayed rather than immediate. Second, planned increases in the saving rate are linked to pay raises. This is meant to diminish the effect of loss aversion – the tendency to weigh losses larger than gains. Because the increase in the savings rate is just a portion of the pay raise, employees do not see their pay fall. Third, once employees sign up for the plan they remain in it until they reach a preset limit or choose to opt out. This uses inertia to keep people in the system. (Thaler and Benartzi 2013: 1152).

Note that for each of the component interventions, Thaler and Benartzi point to a possible causal explanation (none of which involves incentive change – so this also is commonly considered a nudge policy). Yet they do not provide any evidence that these possible explanations are indeed the correct ones. Instead, the evidence they provide is reports of field observations, specifically companies that implemented *Save More Tomorrow*$^{\text{TM}}$. There, employees who chose to join the plan ended up almost quadrupling their saving rate from 3.5% to 13.6% in less than 4 years (Thaler and Benartzi 2004). This evidence shows *how much* the policy intervention, in a particular environment, makes a difference to saving behaviour; but it does not show *how* – through what processes or mechanisms – the intervention produces this behaviour.

Let me focus on the first component intervention of *Save More Tomorrow*$^{\text{TM}}$ for a moment. By asking the employees how much they are willing to contribute at *some time in the future* (rather than how much they would be willing to contribute now), the policy maker is able to get employees to choose larger contribution rates. Thaler and Benartzi here rely on the purported fact of hyperbolic discounting, which comprises two claims (for an overview, see Frederick *et al.* 2002). First, empirical studies indicate that *ceteris paribus* humans and animals place less weight on the future than on the present: they act as though they *discount* future payoffs. For the context of retirement savings, this implies that employees will place more weight on their welfare now than on their welfare in retirement – for them to give up a certain amount of money now to save it for retirement, they must therefore think that this amount will yield them more welfare in retirement than it would yield them now. Second, the discount rates are steepest in the temporal proximity of the respective payoff's realization. This second claim implies a preference reversal: while I prefer one apple tomorrow to two apples next week – because the two apples are heavily affected by the steep discounting that happens between tomorrow and next week – I might prefer two apples in a year to one apple in a year minus one week – because then both items are almost equally discounted, so that the benefit of the extra apple will outweigh the extra discounting of the extra week of delay. For the context of retirement savings, this implies that some employees who would not choose to contribute more to their savings account if comparing 'spending the money now' to 'saving it for retirement' do choose higher contribution rates when comparing 'spending the money in six months time' to 'saving it for retirement'. This is what the *Save More Tomorrow*™ intervention seeks to exploit. Note, however, that hyperbolic discounting here simply summarizes behavioural effects of changing the temporal distance of options – it remains silent on how these changes are produced.

The evidence provided for these policies is of a particular kind, showing that the putative causal factor makes a difference to the putative effect in a specific environment. Such evidence might be produced in controlled experiments (either in the laboratory or the field) – for example in Galilean experiments, in experiments where all background causes are held fixed to one set of values, or in randomized controlled trials (RCTs). Each of these experiments allows measuring the average effect of the intervention on the experimental population and comparing it with the average effect on a relevantly similar population without the intervention. Alternatively, such evidence might be produced through econometric models, which estimate the average effect of the intervention variable by statistically isolating it from other contributing variables in an observational study (Cartwright 2009b; Clarke *et al.* 2014).

Because this evidence is always tied to the environment in which the experiment or observational study was performed, one needs to distinguish efficacy from effectiveness. *Efficacy* is 'the ability of a treatment to produce benefit if applied ideally', while *effectiveness* is 'the benefit that actually occurs when a treatment is used in practice', where 'ideally' is to be understood as a controlled experimental setting (Andrews 1999, cited in Cartwright 2009a: 187–8). I will come back to this distinction in section 4. All I want to note at this point is that such a distinction is important, because much of the evidence for behavioural policy originates from environments different from those in which the policy intervention is supposed to be used in practice – including many field experiments (Cartwright and Hardie 2012). To bridge this distinction, I argue, one must refer to mechanisms – and these mechanisms must be sufficiently evidenced.

What is a mechanism? Although there is no agreement in the literature, there are certain elements that the main contenders all share (Illari and Williamson 2012). In particular, that a mechanism of a phenomenon consists of entities and activities, which are related in some organized fashion, and which are responsible for the production of that phenomenon. Although there might be a unique real mechanism 'out there', when I speak of mechanisms here, I mean *models of a mechanism*, which typically idealize some aspects and generally are abstract representations (or 'mechanism schemes', cf. Craver 2006) of a mechanism. With respect to defaults and hyperbolic discounting, such mechanistic models would describe entities and operations through which the default setting and the change in temporal distance produces the respective behaviour of the targeted individuals. These entities and operations could be described in mental or neurophysiological[2] terms, and some of these (but presumably not all) might also be described in social terms.

In fact, Thaler and Benartzi mention a few candidate mechanistic models, specifically self-control, loss-aversion and inertia.[3] What Thaler and Benartzi do not mention is that these are only some candidate mechanisms for these interventions in these circumstances, and that many alternative candidate mechanisms are discussed in the literature. Thus, the problem is not that there are no mechanistic models associated with these policy proposals, but rather that there are too many – and that there is hardly ever any evidence provided to choose between them.[4] In the next section, I will discuss some of these alternative candidate mechanisms both for default setting and for the *Save More Tomorrow*™ intervention. Then, in section 4, I show that the choice of the mechanistic model has important consequences: it affects our judgements about the effectiveness, robustness, persistence and welfare properties of the respective policy intervention.

3. POSSIBLE POLICY MECHANISMS

Many distinct mechanistic models have been offered as possible explanations of the default effect. I discuss three here.

3.1. Cognitive Effort

The first model proposes *cognitive effort* avoidance as the cause for sticking with the default. As Johnson and Goldstein argue,

[2] For the possible contributions from neuroscience, see Felsen and Reiner (2015).

[3] Not everybody agrees that these three actually amount to mechanistic models (for a critical discussion, see Berg and Gigerenzer 2010). However, for the purposes of this paper, I will assume them to be so.

[4] Mechanistic evidence is evidence for a different claim–i.e. for mechanistic hypotheses–than evidence for the difference-making claims about intervention on the behaviour. This does not mean, however, that mechanistic evidence and difference-making evidence must be two different kinds of evidence (Illari 2011). To the contrary, the same piece of evidence might function both as mechanistic evidence and difference-making evidence.

> making a decision often involves effort, whereas accepting the default is effortless. Many people would rather avoid making an active decision about donation, because it can be unpleasant and stressful. (Johnson and Goldstein 2003: 1338)

Making an 'active decision' involves retrieving information about the respective choice alternatives and deciding upon a strategy how to compare and weigh them. It also involves taking responsibility for the action taken. Their judging a decision as too effortful explains why people tend to stick with the default option. However, this requires that these people are uncertain about their preferences about the choice options (hence the need to make an effort). Someone who already has fully formulated her preferences over the relevant choice options is unlikely to rely on the default instead.[5]

3.2 Loss Aversion

Another mechanism for the default effect describes the systematic influence of the default as a reference point for the evaluation of choice alternatives. In many situations, choice between options involves some trade-off between different dimensions. It has been shown

> that the impact of a difference on a dimension is generally greater when that difference is evaluated as a loss than when the same difference is evaluated as a gain. (Tversky and Kahneman 1991: 1040)

In the context of retirement savings, the relevant dimensions might be 'current consumption' and 'financial security after retirement'. To choose which amount to save thus implies accepting a trade-off between differences in these two dimensions: saving a lot yields higher future security but lower current consumption than saving little – and vice versa. Rational choosers determine how much weight they put on having more current consumption or more financial security, and form their decisions accordingly.

According to the loss aversion model, however, the weight assigned to the differences in these dimensions depends systematically on whether the chooser sees them as a gain or a loss: losses generally are weighed higher than gains. Whether a difference in a dimension is interpreted as a gain or a loss depends on a reference point, typically the status quo. For example, if the chooser considers reducing consumption from current levels, then she interprets it as a loss; when she considers raising it above current levels, then she interprets it as a gain.

[5] This claim is supported by a recent observational study: 'We find that the likelihood of default is significantly lower for those with higher levels of general and decision-specific knowledge' (Brown *et al.* 2012: 22).

Setting the default, it is claimed, affects what the chooser sees as this 'current level', and hence sets the reference point against which a change is identified either as a gain or a loss. For example, setting a high retirement fund contribution as the default lets the chooser interpret alternative choices with lower contribution rates as a gain in current consumption, and a loss in future financial security. Consequently, according to the loss aversion model, she will put more weight on the financial security under this default, than she would if the default were a low contribution. This different weighing of the same differences explains why people tend to stick to default options: seen from the default, the alternatives are largely evaluated as losses, and hence not preferred to the default itself.

3.3. Recommendation Effect

A third explanation describes the default as a signalling device. The policy maker who sets the default thereby signals (or 'recommends') that she judges the default option best for the choosers. The choosers who receive the signal and trust the policy maker's judgement infer that it is in their interest to choose the default (McKenzie *et al.* 2006; Gigerenzer 2008: 24). Setting the default thus provides information that increases the choosers' motivation, *ceteris paribus*, to choose the default option.

Similar to default setting policies, there are many possible mechanistic models for the effects of policies intervening on the time-horizon. As I discussed in the previous section, these models in particular must clarify why choosers discount utilities hyperbolically. I sketch two such models here.

3.4 Visceral Factors

The first model focuses on time-dependent effects of visceral factors. Such factors include drive states like hunger, thirst and sexual desire, cravings caused by drug addiction, as well as moods, emotions and physical pain. As influences on behaviour, they are typically distinguished from motivations based on perceived self-interest. The influence of these factors is not always present, but rather is triggered by certain cues, which include physical closeness, sensory contact (e.g. the sight, smell or feeling of a craved object), as well as temporal proximity to obtaining the craved object. Loewenstein describes this time-dependent influence on behaviour thus:

> Visceral factors operating on us in the here and now have a disproportionate impact on our behavior. Visceral factors operating in the past or future, or experienced by another individual are, if anything, underweighted. (Loewenstein 1996: 276)

Consequently, while agents expect that certain objects in the future will make them hungry, sexually aroused or irate, they nevertheless evaluate these future objects according to their perceptions of self-interest, and more or less free from the influence of these visceral factors. They therefore often prefer acting against their expected hunger, arousal or ire. Yet when they come closer in time to experiencing these objects, the visceral factors increasingly influence their valuation, so that when the time comes to choosing between these objects, hunger, arousal or ire often has become the determining factor of their behaviour. This explains the hyperbolic discounting of impulse-inducing objects: the influence strength of visceral factors on the individual's evaluation increases proportionally with temporal proximity to the present.

For the context of retirement savings, asking the employee how much they are willing to give up *now*, in order to save for retirement much later, exposes the evaluation of loss in current consumption to (potentially strong) visceral factors, while leaving the evaluation of later financial security to (potentially less strong) perceived self-interest considerations. The employee is therefore more likely to reject additional savings contributions. Asking the employee instead how much they are willing to give up in, say, six months time, in exchange for an increased retirement fund later, will expose the evaluations of both prospects to perceived self-interest considerations, and thus more likely will yield higher contribution rates.

3.5 Uncertainty

A second model, coming out of biology, identifies the difference in uncertainties between the nearer and the more distant options as the cause for hyperbolic discounting. Temporal distance is associated with uncertainty whether the chosen option will be realized at all: 'I perceive a promised future reward not as a sure thing, but instead as having a probability attached to it' (Sozou 1998: 2017). In biology, the uncertainty is typically that of dying before the animal can consume the promised future reward. If this hazard rate is constant for each time step, then an exponential discounting function results. But Sozou argues that for many animals, particularly those that experience heterogeneous ecological environments, these hazard rates will not be constant. Furthermore, if there is uncertainty about the hazard rate, animals might have a probability distribution over it, and for many such distributions, a non-exponential discounting rate for future rewards results. More specifically, if the hazard rate is characterized by an exponential distribution, Sozou shows that the hyperbolic discounting function ensues (Sozou 1998: 2017).

Consequently, this model explains behaviour in accordance with the hyperbolic discounting function as based on considerations that take seriously the uncertainty of getting a temporally distant reward. For the context of retirement savings, this implies that employees considering giving up some of their current consumption for more pension savings later trade a certain outcome for an uncertain one; while those considering giving up some of their consumption in six months time for more pension savings trade one uncertain outcome for another one. Given that this uncertainty is characterized by specific distributions, the relevance of this certainty might be so high that employees choose higher contributions in the second than in the first case.

This list is not meant to be complete, but only to illustrate the differences between the mechanisms on which these possible explanations rely. They are different in the sense that it is 'possible to interfere with the operation of one without interfering with the operation of the other' (Woodward 2003: 48). For example, with respect to defaults, a reduction in preference uncertainty would affect the operation of the cognitive effort mechanism, while leaving the operation of the recommendation effect unchanged. Similarly, with respect to hyperbolic discounting, a change in the intensity of emotions and visceral effects (for example through medication) would affect the operation of the visceral factor mechanism, while leaving unchanged the uncertainty mechanism. Of course, that these mechanisms are different from each other does not exclude that they operate side-to-side in a population or even in an individual.[6]

Proponents of behavioural policies, while sometimes discussing these possible explanations, typically do not provide any evidence for them. Nor do they provide any evidence for the distribution of such mechanisms in the populations or individuals on which their policies intervene. Instead, they seem content with merely difference-making evidence for their policies, adding – if at all – the mechanistic models as a suggestion 'how it could be' (cf. e.g. Johnson and Goldstein 2003; Thaler and Benartzi 2004; Beshears *et al*. 2009). In the next section, I show that this deficit threatens to undermine the justification of such behavioural policies, as these policies' effectiveness, robustness, persistence and welfare effects are sensitive to the question which mechanisms are the correct ones for the intended target context.

4. THE NEED FOR MECHANISTIC EVIDENCE

Perhaps it is tempting to think that difference-making evidence is enough for policy purposes: policies, after all, are for making the right kind of difference in the world. Perhaps, then, mechanistic evidence is needed just

[6] There is some evidence for such population heterogeneity: '[O]n why people default [we] find … a surprising degree of heterogeneity, with no more than about one-third of employees citing any one particular reason for defaulting' (Brown *et al*. 2012: 22).

for *other* purposes, like better understanding, but not for justifying a policy itself? In this section, I show that this is not so.

My argument proceeds by showing that important aspects of a policy intervention are sensitive to the underlying mechanistic assumptions. Specifically, I show that in certain environments, one and the same policy is judged effective, robust, persistent or welfare-improving when assuming one kind of mechanistic model, while it is judged not so when assuming another mechanistic model.

4.1 Effectiveness

Let's assume that we have good evidence that a certain policy is efficacious – i.e. that its intervention makes a difference to the variable of interest under some ideal experimental conditions. But what matters for the justification of this implementation is of course that the policy is effective in the target environment and population – i.e. that the policy intervention makes a difference to the variable of interest under the actual conditions in which it is supposed to be applied. This difficulty is of course well known. See for example Cartwright (2009b: 133):

> Efficacy is no evidence whatsoever for effectiveness unless and until a huge body of additional evidence can be produced to show that efficacy can travel, both to the new population and to the new methods of implementation.

One strategy to overcome this external validity problem is to summon mechanistic models as arguments for why efficacy can travel in this way. In this argument I follow Steel (2008), who maintains that mechanistic evidence indicates how an intervention works in the ideal experimental situation, *and* to what extent the same mechanisms are also operative in the target population. The latter information then allows us to judge whether an efficacious intervention is likely to be effective in the target population. Specifically, this inference depends in at least two ways on mechanistic information: first, whether the necessary background conditions are in place, and second, whether the policy is implemented in the right way.

A policy intervention will be efficacious in a certain environment, only if all *necessary background conditions* are in place. To know which conditions to check for in the target environment requires mechanistic information. Take the default-setting example. Under the cognitive effort explanation, default setting is efficacious only if people's deliberation costs are sufficiently high. Without this condition in place, the intervention simply won't work. Thus we must check whether it is satisfied in the target environment. If in the target environment people do have considerably lower deliberation costs than in the original case described by Beshears *et al.* (2009), then we should expect, *ceteris paribus*, that setting a new default would not be effective in the target environment.

Under the recommendation effect explanation, in contrast, differences in deliberation costs would not matter: either way, the default would provide new information that would affect choice. Thus, identifying necessary background conditions for the policy's efficacy in the target environment depends on mechanistic information.

Furthermore, a policy intervention will be efficacious in a certain environment, only if the intervention is *implemented* in the right way. In an ideal intervention, the intervention acts as a 'switch', interrupting the causal influence of all other causal factors on the variable intervened on (Woodward 2003: 102). But depending on the assumed mechanism, not all proposed behavioural interventions satisfy this ideal.

Take the *Save More Tomorrow*™ case, under the uncertainty explanation. Without the proposed intervention, the temporal difference between the two rewards impacts the differential uncertainty judgement, which in turn influences choice. But temporal difference is of course not the only influence on differential uncertainty judgements: other information – e.g. about institutional stability – also matters. An intervention that changes the temporal difference between the rewards, as proposed by Thaler and Benartzi (2004), acts as a switch only on temporal difference but not necessarily on all the other influences on uncertainty. For example, a clumsy communication and implementation of the *Save More Tomorrow*™ plan might also create the impression that retirement plans are changed haphazardly, letting employees revise their uncertainty judgments about their ability to retrieve funds. Such an increase in uncertainty, caused by the implementation, might well erase any positive effects that the policy would otherwise have had on contribution rates.

Other mechanisms, in contrast, do not exhibit the same implementation-sensitivity. Under the visceral factors explanation, for example, how the time-horizon change is communicated will *ceteris paribus* make no difference. Thus, whether and how a policy is implementation-sensitive depends on the underlying mechanism.

More generally, externally valid inferences require mechanistic evidence. Some have argued that the issue of external validity can be avoided altogether to the extent that difference-making evidence is produced by 'natural field experiments' in the target environments itself (Levitt and List 2009). However, there are considerable time differences between field experiments and the implementation of the policies in that same population.[7] Because during such time periods,

[7] The implementation of Thaler and Benartzi's proposal, for example, although a widely celebrated example of behavioural policy, is still very much under debate in 2015 – eleven

a population typically undergoes a lot of changes (e.g. demographic, political, technological), the external validity issue and the consequent need for mechanistic evidence arises here too. Furthermore, there are particular aspects of effectiveness that cannot be dealt with in this way. To those I turn now.

4.2 Robustness

Effectiveness concerns 'the benefit that *actually occurs* when a treatment is used in practice' (see section 2). However, this lumps a lot of things together. Effectiveness is supposed to be inferred from efficacy (plus additional information, as argued in 4.1). But efficacy, as I use it here, is evidence for the intervention making a difference for the variable of interest in an ideal environment. At least under the approach proposed here, efficacy does not cover welfare judgements as implied in 'benefit', nor other effects besides the variable of interest, as implied in 'that actually occurs'. In order to capture more clearly these features – what I call robustness, persistence and welfare effects – I will now treat them as separate aspects of effectiveness.

Take robustness first. An intervention I intervenes on a causal variable X, causing a change in the effect of interest Y. I call such an intervention robust if 'any direct path from I to Y goes through X' (Woodward 2003: 98). This excludes non-robust cases where the policy intervention also affects secondary factors, which in turn influence the variables of interest, so that the total effect on Y might be cancelled or even reversed. Yet social policies often exhibit such complications, as argued e.g. by the Lucas Critique (see Cartwright 2012 for further examples and discussion).

Let me illustrate this with an example. Because doctors who experience a conflict of interest tend to treat patients differently than without such a conflict, policies have been proposed that would make the disclosure of such conflicts of interest mandatory. However, Sah, Cain and Loewenstein in a number of recent papers have pointed out that such disclosures might have perverse effects. Specifically, the disclosure in the target population makes a difference in that it successfully informs people's deliberations about these conflicts of interest (and hence is effective in this sense). Yet this effectiveness also triggers two other processes that lead to *increased*, instead of the expected *reduced* compliance (Sah, Loewenstein and Cain 2013). In particular, the *insinuation anxiety* lets advisees fear that rejecting advice may signal to the advisor that they believe the advisor is corrupt; and the *panhandler effect* lets advisees

years after its first publication. See Thaler and Benartzi (2013) and Benartzi and Lewin (2012).

feel the pressure to help advisers obtain their personal interests once the adviser discloses this interest (Cain, Loewenstein and Moore 2011). Note the difference to the effectiveness issue of implementation above. There, the variable of interest was not affected by the intervention, because alternative effects cancelled each other out. Here, the variable of interest is affected (i.e. the patient is informed by the disclosure), but the intervention also causes other variables relevant for 'what actually occurs' (i.e. the disclosure also influences the patient's deliberation in a way that counteracts the influence of the disclosed information).

Whether a policy suffers from non-robustness depends on the underlying mechanism – and how to look for such complications depends on the mechanistic information we have. For example, a policy that aims at making people exercise more by default subscribes all employees to ten monthly gym visits. Let's assume that the policy makes a difference: after it is implemented, people spend more time in the company gym than before. But that is not the ultimate desired effect of this policy – rather, making people exercise more is. But under the cognitive effort explanation, there might be environments in which the intervention does not produce this effect. For example, people might have ended up more often at the company gym, because making an active choice which gym to go to was too effortful. Yet they still might keep track of their overall exercise time, and cut back proportionally on their previous exercise time outside of the company gym. After all, the default change has not convinced them that exercise is (more) important for them than before – it has only made them end up with more time at the company gym due to their indecisiveness.

Such a compensation or crowding-out effect is quite plausible under the cognitive effort explanation, at least for some contexts. Before implementing such a policy, the policy maker should have investigated whether such a non-robustness threatens the effectiveness of the planned intervention.

Under the recommendation explanation, in contrast, no such fragility is plausible. If the default setting yields a signal of sufficient quality, and if the decision maker did not possess the full information before, then the default setting will influence his evaluation and choice, *ceteris paribus* leading to an overall increase in exercise time.

Thus, our judgement on whether a policy is robust depends on mechanistic information. Furthermore, such complications are unlikely to be picked up by field experiments in target environments, as these are designed to test an intervention's efficacy. Of course, one could design field experiments to identify non-robustness (taking into account the relevant variables of interest and allowing for much longer time frames). But to design them, one needs to know what complications to look for – and this already requires mechanistic information.

4.3 Persistence

Another kind of complication arises from interventions that somehow affect the causal relation between the variable X intervened on and the effect of interest Y. Such a 'structure-altering intervention' (Steel 2009: 155), if it immediately cuts or changes the relation between X and Y, is simply a case of non-effectiveness in the target environment, as discussed in 4.1. However – depending on the mechanism – the intervention might also be (initially) effective and robust in a target environment, yet later the accumulated effect of its repeated implementation has this structure-altering consequence. Policies that do not suffer from this complication I call *persistent*.

A typical example of non-persistence is a wearing off of the intervention's effect in the longer run. I am not aware of rigorous studies of wearing off of nudges, but such effects have been observed in incentive-altering interventions, and are plausible for nudges as well.[8] When such a policy is first implemented, it produces the desired effect. But by repeated implementation, this causal connection eventually weakens or breaks completely. Whether this is likely to happen again depends on the mechanism that connects intervention and the variable of interest.

For example, consider again the gym-exercise example on the assumption that the default works because of the cognitive effort mechanism. One effect of increased gym attendance will be that with time, employees will have better-informed preferences about the gym. That is, they will discover whether they like going to the gym and whether they feel it is worth their time. As argued in 3.1, however, that people grow more certain about their preferences implies that cognitive effort will have less of an effect on them. In that case, an initial increase in the use of the gym as a result of the new default might be followed by a drop-off, as only those who actually like going to the gym stick with it.

Such a wearing-off effect is not plausible under the assumption that the default works because of the loss-aversion mechanism. To the contrary, one should expect that the more people accept the default option as the status quo (which is likely to happen with the passing of time) the more they will consider switching away from it as a loss of exercise time, and the more weight they will put on this loss-dimension (rather than the dimension of gaining more free time for other activities, etc.). Thus, our judgement on whether a policy is persistent depends on mechanistic information.

[8] For example, Strickland (2014) reports of a yet to be published review of 39 studies, most assessing smoking, which concludes that although incentives did increase behaviour change, their role was potentially limited because the effects wear off three months after the incentive ceases.

Non-persistence is unlikely to be picked up by field experiments on target populations, as this would require very broad and long study perspectives. Instead, indications that such factors might be at work are better obtained from experiments that explicitly seek to produce mechanistic evidence.

4.4 Welfare

Effectiveness not only concerns the causal impact on a variable of interest, but also 'the *benefit* that actually occurs when a treatment is used in practice' (see section 2). Typically, when comparing different policies, the policy maker first identifies a relevant evaluation criterion, and then evaluates the respective expected policy effects according to this criterion.

Proponents of nudge fit this general pattern: they evaluate different policy options according to a subjective welfare criterion. Their self-declared goal is to 'steer people's choices in welfare-promoting directions' (Sunstein and Thaler 2003: 1159). More specifically, they defend an informed-preference account of subjective welfare, which takes individual preferences as welfare-relevant if they are sufficiently well informed and result from deliberations that are taken in a 'cool' and alert state.[9] Nevertheless, they insist on a consequentialist perspective, which evaluates only the effects of policy interventions, but disregards the processes through which these effects are produced (see e.g. Sunstein and Thaler 2003: fn 22). From this perspective, mechanistic evidence has no influence on the welfare judgements themselves.

In contrast to this view, I argue that to assess the welfare effect of an effective behavioural policy often requires assessing this result in the light of how it was produced. The basic idea is that nudges produce behavioural effects through varying degrees of manipulation, that manipulation 'perverts the way that a person reaches decisions, forms preferences or adopts goals' (Raz 1986: 377–8), and that such perversely formed preferences do not satisfy the well-informed criterion of welfare that e.g. Sunstein and Thaler subscribe to. In order to determine how manipulative a nudge in a particular target population is, one must investigate the mechanism through which the intervention produces the behaviour. Consequently, welfare effects of nudge policies depend on the mechanisms that drive them.

Nudge proponents occasionally acknowledge the relevance of such manipulating mechanisms. Johnson and Goldstein, for example, concede

[9] Their reliance on the informed-preference account becomes particularly clear when they justify nudge interventions on cognitively biased behaviour: 'In some cases individuals make inferior decisions in terms of their own welfare – decisions that they would change if they had complete information, unlimited cognitive abilities, and no lack of self-control' (Sunstein and Thaler 2003: 1162).

that some defaults might induce people to 'become donors against their wishes' (Johnson and Goldstein 2003: 1339), and furthermore that some default mechanisms are more manipulative and hence more problematic with respect to welfare than others: 'one might draw different [normative] conclusions if the effect of defaults on donation rates is due primarily to the physical costs of responding, than if they were due to loss aversion' (Johnson and Goldstein 2003: 1339).

To illustrate such a case, let's imagine a default setting policy that aims at increasing employees' retirement contributions, and that is effective because of the recommendation effect mechanism. That is, employees take the policy maker's setting of the default as a relevant piece of information about how much retirees on average need to continue a good life, and incorporate this information in their deliberation. To the extent that the policy maker did not set the default to intentionally deceive, the employees' decisions will not be manipulated – but rather informed – by the default.

In contrast, now imagine the same situation but driven instead by the cognitive effort mechanism. For example, people are too lazy to inform themselves and determine their own preferences for how much they want to have later and how much they want now. Instead, they go with the default contribution rate. Consequently, employees are manipulated by the default, which affects their preferences without engaging their deliberative faculties.[10] Hence, whether two instances of default-setting, which produce the same behaviour, have the same welfare consequences or not depends on which mechanisms produce it.

Nudges not only produce effects through exploiting mechanisms of differing degrees of manipulation. Also, an important argument *for* nudges is that they improve people's welfare by helping them to avoid manipulative forces. These cases are the inverse of those discussed above: here a manipulative status quo is improved upon by a nudge policy

[10] That defaults affecting choices through cognitive effort are more manipulative than those affecting choices through the recommendation effect is supported by the following study of choosers' subsequent regrets. The study investigated a large public retirement system that explored employees' reasons to default and whether they subsequently regret their decision (Brown *et al*. 2012). It found that reasons like being 'too busy', 'lack[ing] of information' or the choice being 'too complex' increased the probability of employees choosing the default. This is evidence for some defaults producing the effect through the cognitive effort mechanism. It also showed that the same people who chose the default for these reasons were more likely to regret their decision later: between 55% and 75% of those defaulting regretted or strongly regretted being in the default plan (Brown *et al.* 2012: 23). In contrast, of those who chose the default because of the recommendation effect, only 43% regretted or strongly regretted being in the default plan. To the extent that people are more likely to regret a manipulated choice, this provides evidence that the cognitive effort mechanism is more manipulative than the recommendation effect.

that encourages people to choose more in accordance with their true preferences, and hence can be expected to be welfare-improving. I argue that the validity of this argument again depends on the mechanisms assumed to produce the behaviour.

Consider a case of time-horizon shifts. ‘The visceral factor perspective’, Loewenstein (1996: 289) writes, ‘helps to explain when and why people view their own, and others’, behavior as irrational’. The evaluation of present rewards is determined by visceral forces, which exclude relevant information and contradict non-emotional preferences. It is therefore not welfare relevant. From that perspective, shifting the time horizon helps people avoid ‘behaving contrary to [their] own perceived self-interest’ (Loewenstein 1996: 289): it helps free people’s choices from visceral influences, and instead allows people to choose so as to satisfy their consistent and well-informed preferences. Consequently, the changed behaviour brought on by a time-horizon shift, under the visceral forces explanation, constitutes a welfare improvement.

In contrast, Sozou (1998: 2017) insists that ‘There is . . . no inconsistency if I perceive a promised future reward not as a sure thing, but instead as having a probability attached to it’. Consequently, judging a sooner, smaller reward as less uncertain and hence as preferable to a larger later one, is compatible with the consistency and well-informed criteria. If an agent holds justified beliefs about the distribution of the hazard rate consistent with hyperbolic discounting, then her choices in accordance with hyperbolic discounting are as welfare-relevant as her evaluations of larger, later rewards. From that perspective, policies that change people’s choices through time-horizon shifts *ceteris paribus* do not constitute a welfare improvement. Thus, our judgement of whether a policy effect constitutes a welfare improvement depends on mechanistic information.

How should one incorporate such welfare-relevant mechanistic information in the overall policy decision process? I suggest that this should proceed in analogy to the cases 4.1–4.3. There, mechanistic evidence functions as a filter for inferences from experiments to the target population. Analogously, mechanistic information should function as a filter of welfare judgements of policies’ consequences. For example, if a policy produces effects that appear as welfare improving in comparison to the status quo, then the discovery that the policy operates through a manipulating mechanism constitutes a reason to reject this comparative judgement. Furthermore, if the mechanistic evidence allows quantitative inferences (Brown *et al*. 2012 for example differentiate the population by ‘reasons for defaulting’) – e.g. that under policy intervention, 40% of the behaviour is produced by a manipulative and 60% by a non-manipulative mechanism – then this information can be used to weigh the purported welfare effect.

5. CONCLUSION

Mechanistic information is often highly important for assessing the effectiveness and welfare consequences of a given policy. In particular, without the right kind of information, we often cannot tell whether in a particular context a policy is effective, is robust or persistent, nor whether it has a positive welfare effect. Yet these properties are often crucial for the justification of the policy in this context: without efficacy, we cannot justify the inductive inference from a study environment to the target environment. Without robustness and persistence, we often cannot justify the policy even from difference-making evidence obtained from the target environment itself. And without welfare-relevance, we do not know whether the policy affects preference in accord with the assumed normative principles of welfare-relevance. In other words, without specifying the underlying mechanism, we are often not able to say whether a behavioural policy implemented in a specific environment is justified by given difference-making evidence or not.

This result has at least three implications. The first is practical – it cautions against a premature implementation of the proposed policies. Most of the behavioural policies currently proposed do not provide any mechanistic evidence. Yet without such mechanistic information, as the cases discussed here show, one often cannot justify these policies.

Arguably, it is very difficult to obtain evidence about mechanisms in the target environments that fully justifies externally valid inferences (Steel 2008: 166). Instead, we often have to make do with less and with lower-quality mechanistic evidence. By this I mean evidence that does not support a precise mechanistic model of the target situation, but that provides only qualitative information regarding factors upon which the policy's effectiveness and welfare-properties is likely to depend (cf. Steel 2009: 161). Searching for reasons to think that a policy in its target environment is not efficacious, robust, persistent or welfare-improving might yield no result – which in itself would contribute to the justification of that policy. But search one must. Thus I propose the following *sufficiency principle*: a policy is based on sufficient mechanistic evidence if it takes all available mechanistic evidence into account, where availability is constrained by current theoretical and technological feasibility. If information of this sort does not enter the discussion at all, these policies cannot and should not be described as 'evidence-based'.

The second implication is philosophical – it demonstrates the importance of distinguishing different kinds of causal information and their respective relevance in justifying policy. In the first place, I demonstrated that beyond difference-making evidence we also need mechanistic evidence. Because this is evidence about different claims (not different kinds of evidence), the one cannot be substituted for the other – simply providing better difference-making evidence does not compensate for a lack of mechanistic evidence. In the second place, I argued that mechanistic evidence should be further differentiated: necessary background conditions, implementation sensitivity, robustness, persistence and welfare-relevance require different kinds of causal information. This list is probably not complete. Nevertheless, by distinguishing between different kinds of causal information and showing their relevance for specific purposes, it goes beyond the general claim that without mechanisms, no social policy can be justified. Furthermore, this distinction helps structure the search for mechanistic evidence that is maximally relevant for the justification of a concrete policy in a specific environment.

The third implication is methodological – it raises the question which kind of empirical investigations provide this information. Early behavioural policies often were supported by difference-making evidence from laboratory experiments. Recently, it has been argued that some of the concerns about the effectiveness of the proposed policies can be resolved by relying on field experiments instead (Levitt and List 2009). This might be true for some of the concerns raised in this paper, for example the issue of implementation sensitivity. But other worries remain, especially those concerning persistence and welfare properties. Field experiments typically do not last long enough to exclude the existence of detrimental feedback effects. And because they typically do not reveal the underlying mechanism, one cannot judge on their basis whether the policy affects preference in accord with the assumed normative principles of welfare-relevance. Thus, this paper not only cautions against premature policy implementation and argues for further differentiation of causal information, but it also provides an argument why field experiments – by themselves, without mechanistic information – are not a sufficient evidential source for evidence-based policies.

ACKNOWLEDGEMENTS

In 2014, earlier versions of this paper were presented at the TINT seminar in Helsinki and at the PSA meeting in Chicago. I am indebted to the audiences there for their insightful and constructive comments. This research was supported by Swedish Research Council grant no. 2011-1302 as part of the project 'Assisted Decision Making'.

REFERENCES

Andrews, G. 1999. Efficacy, effectiveness and efficiency in mental health service delivery. *Australian and New Zealand Journal of Psychiatry* 33: 316–322.

Benartzi, S. and R. Lewin. 2012. *Save More Tomorrow: Practical Behavioral Finance Solutions to Improve 401(k) Plans*. New York, NY: Penguin.

Berg, N. and G. Gigerenzer. 2010. As-if behavioral economics: neoclassical economics in disguise? *History of Economic Ideas* 18: 133–166.

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

Brown, J. R., A. M. Farrell and S. J. Weisbenner. 2012. *The Downside of Defaults* (No. orrc12-05). National Bureau of Economic Research. http://www.nber.org/aging/rrc/papers/orrc12-05.pdf. (Accessed 22 March 2015)

Cain, D. M., G. Loewenstein and D. A. Moore. 2011. When sunlight fails to disinfect: Understanding the perverse effects of disclosing conflicts of interest. *Journal of Consumer Research* 37: 836–857.

Cartwright, N. 2009a. What is this thing called 'efficacy'? In *Philosophy of the Social Sciences. Philosophical Theory and Scientific Practice*, ed. C. Mantzavinos, 185–206. Cambridge: Cambridge University Press.

Cartwright, N. 2009b. Evidence-based policy: what's to be done about relevance? *Philosophical Studies* 143: 127–136.

Cartwright, N. 2012. Will this policy work for you? predicting effectiveness better: how philosophy helps. *Philosophy of Science* 79: 973–989.

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

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

Craver, C. F. 2006. What mechanistic models explain. *Synthese* 153: 355–376.

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

Frederick, S., G. Loewenstein and T. O'Donoghue. 2002. Time discounting and time preference: a critical review. *Journal of Economic Literature* 40: 351–401.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science* 3: 20–29.

Grüne-Yanoff, T. and R. Hertwig. 2015. Nudge versus boost: how coherent are policy and theory? *Minds and Machines*. doi: 10.1007/s11023-015-9367-9.

Halpern, D. 2014. Nudge unit: our quiet revolution is putting evidence at heart of government. *The Guardian*, 4.2.2014. http://www.theguardian.com/public-leaders-network/small-business-blog/2014/feb/03/nudge-unit-quiet-revolution-evidence. (Retrieved 26 February 2014).

Illari, P. M. 2011. Mechanistic evidence: disambiguating the Russo–Williamson thesis. *International Studies in the Philosophy of Science* 25: 139–157.

Illari, P. M. and J. Williamson. 2012. What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science* 2: 119–135.

Johnson, E. and D. Goldstein. 2003. Do defaults save lives? *Science* 302: 1338–1339.

Levitt, S. D. and J. A. List. 2009. Field experiments in economics: the past, the present, and the future. *European Economic Review* 53: 1–18.

Loewenstein, G. 1996. Out of control: visceral influences on behavior. *Organizational Behavior and Human Decision Processes* 65: 272–292.

McKenzie, C. R., M. J. Liersch and S. R. Finkelstein. 2006. Recommendations implicit in policy defaults. *Psychological Science* 17: 414–420.

Park, C. W., S. Y. Jun and D. J. MacInnis. 2000. Choosing what I want versus rejecting what I do not want: an application of decision framing to product option choice decisions. *Journal of Marketing Research* 37: 187–202.

Raz, J. 1986. *The Morality of Freedom*. Oxford: Oxford University Press.

Sah, S., G. Loewenstein and D. M. Cain. 2013. The burden of disclosure: increased compliance with distrusted advice. *Journal of Personality and Social Psychology* 104: 289–304.