Ethics of nudges: a general framework with a focus on shared preference justifications

Christine Clavien
University of Geneva

Abstract

Nudges are soft interventions designed to influence human behaviour. They are increasingly used by policy-makers around the world and generate heated controversies over their ethical acceptability. In this paper, I depict nudges as science-based technologies, point to new ethical responsibilities they raise – when they are used but also when they are not used – and to the necessity of establishing clear assessment criteria. I then elaborate a general framework for assessing the acceptability of particular nudges. The second part of the paper focuses specifically on “shared preference justifications”, that is, arguments of the sort “nudges help people make better choices as judged by themselves.” I explain why constructing shared preference justifications is important but difficult, and propose a framework for doing so. This framework takes autonomy concerns seriously without overstating them.

Keywords: nudge, ethics, moral, autonomy, shared preference justification

1. Introduction

1.1. What are nudges

Nudges are soft interventions produced by “nudgers”, who may be policy-makers – e.g. states –, decision boards of organizations and institutions – e.g. private companies, hospitals – or more generally individuals in positions of power. These interventions introduce small modifications in the choice architecture of target populations, the “nudgees”, and thereby steer their behaviour in predictable directions. To qualify as nudge, however, these interventions should not significantly change material incentives and should leave all previously available choice options open (Thaler & Sunstein, 2008). This is illustrated in Figure 1 (p.8) with the example of policy-makers who replace ordinary grey staircases by luminous stairs, with the aim of getting citizens to make exercise instead of taking elevators. Before and after the nudge is applied, citizens’ space of choice options – either climbing the stairs or taking elevators – remains basically the same.

Most nudges operate by exploiting unconscious decision-making pathways, such as risk aversion (Tversky & Kahneman, 1991), compliance to authority (Milgram, 1965), default (S. D. Halpern, Ubel, & Asch, 2007), hyperbolic discounting (Frederick, Loewenstein, & O’Donoghue, 2002), compliance with social norms (Cialdini & Goldstein, 2004), or prejudices against some categories of individuals (Chapman, Kaatz, & Carnes, 2013). Nudges may help to bypass, eliminate, or elicit unconscious decision-making pathways in order to induce desirable behavioural choices. For example, the risk aversion bias leads patients to underestimate the gains of a drug compared to its possible side-effects. This reasoning failure may be corrected by framing medical information so as to emphasize the risks associated to non-consumption of the drug (Redelmeier, Rozin, & Kahneman, 1993). In another example, knowing that people are strongly influenced by social norms, informing citizens that their neighbours display energy saving habits will induce an increase in energy saving behaviour (Nolan, Schultz, Cialdini, Goldstein, & Griskevicius, 2008).

There are debates about the correct definition of “nudge” and thus, about what interventions count as nudges, as opposed to mandates, incentives, mere provisions of information, or manipulations (for more on this topic, see Nys & Engelen, 2016). Moreover, in specific cases, it may be debatable
whether a policy meets the desired criteria – e.g. whether it really leaves open the original choice options. These debates are of little importance for this paper since the general framework I propose applies to nudges and soft “nudge-like” interventions in general.

1.2. Nudges are technologies and create new moral responsibilities

Nudges can be seen as science-based technologies which impose new responsibilities. Let us draw a parallel with a medical technology. Once defibrillators have been invented and their usefulness for saving lives at little expense demonstrated, health policy deciders committed to improving citizen’s quality of life have a moral responsibility to take a stance about making defibrillators available to citizens. Suspending their decision is not ethically neutral. They would need good reasons for not making defibrillators available in hospitals – e.g. showing that limited state funding is better spent in other public utility domains. Correspondingly, nudges allow their users – i.e. nudgers – to achieve or access states of affairs that were less easily accessible before. If these opportunities have a social impact that is considered desirable or destructive, they impose a re-evaluation of previously available options, and new moral responsibilities upon nudgers. If there is scientific evidence about a cost-effective nudge with desirable social impact, informed deciders who aim at improving citizens’ quality of life need to have good reasons for not putting it into practice.

Take the example of storm naming. Since 1979 in the US, storms are named alternatively with male or female names. A scientific study (Jung, Shavitt, Viswanathan, & Hilbe, 2014) has provided some evidence that female-named storms cause significantly more deaths than men-named storms. The authors of the study make the explanatory hypothesis that a female name induces a lower perceived risk, thus lower protective actions ahead of the storm. If this analysis is scientifically robust, it indicates that a simple framing effect dramatically impacts on human lives, and State officers could easily nudge citizens’ protective behaviour by naming all future storms with male names. In light of such information, State Officers committed to protect citizens’ security need good reasons for not nudging citizens in this way. One such reason may be a gender equity concern. Another reason may ground on doubts about whether the study meets scientific standards. In this particular case, the study indeed has statistical flaws (Smith, 2016), which is a good reason for maintaining the status quo on storm naming.

Once we recognize that much of our behaviour is influenced and biased by all sorts of unconscious psychological mechanisms, and once we recognize that some nudges may help us avoid risk factors or contribute to some valuable goals, it becomes problematic – at least not ethically neutral – to ignore or reject the possibility to use nudges (Reiss, 2013; Skipper, 2012). Simultaneously, as illustrated with the storm example and as will become clear later in this paper, scientific evidence is particularly relevant for the assessment of nudges.

2. Assessment of nudges – a general framework

Facing the development of nudge technologies and the new responsibilities they create, we can expect that individuals in positions of power will increasingly make use of nudges in all sorts of situations. This possibility raises a series of ethical concerns because, as with any technology, nudges can be used for other-regarding as well as self-regarding reasons, and they can produce unwanted side-effects.

A number of scholars have convincingly argued that ethical concerns related to nudges may arise on occasion, while there is no good reason to think that any nudge inherently and by necessity is ethically problematic (see for instance Ashcroft, 2011; Blumenthal-Barby & Burroughs, 2012; Nys & Engelen, 2016).
2016; Thaler & Sunstein, 2008). I will take this conclusion as granted. What is presently needed is a simple and comprehensive framework that may help nudgers to evaluate the acceptability of a nudge early on in the process of its elaboration. Nudgers need to ask the right questions that will help them identify all reasons in favour and against the nudges they plan to use.

In the following sections, I will sketch the underlying principles of such a “nudge assessment” tool. I will argue that three main questions should be addressed. First (section 2.1), are the goals that the nudge is deemed to reach ethically justifiable? Second (section 2.2), is the nudge an effective means to achieve these goals? Third (section 2.3), are there ethical concerns raised by the application of the nudge? After weighing-up all the pro and contra aspects raised by this tree-level investigation, one can take a balanced decision in favour or against a nudge. Let us consider each question in turn.

2.1. Justification for nudgers’ goals

While designing a nudge, nudgers may pursue three types of goals which are not necessarily exclusive: (a) “Selfish goals” are about increasing nudgers’ own “interests” – i.e. their own benefit, well-being, or good. Examples include nudging for being elected, for selling a product, or for favouring one’s close family or friends; (b) “Social goals” are about increasing a societal interest. Examples include nudging for more social stability, or for reducing discrimination against some minority groups; (c) “Nudgee-driven goals” are about increasing nudgees’ interests. Examples include nudging for improving nudgees’ health or retirement savings. In some cases, social and nudgee-driven goals may be conjointly realizable. Examples include nudging for safer driving, or for taxpaying compliance – assuming that the collected funds are used in the interest of citizens. In other cases, selfish or social goals may be achieved at the expense of nudgees’ interests. Examples include nudging for selling unnecessary or unhealthy products, or for increasing organ or charity donation².

To evaluate the appropriateness of a nudge, the first main question to ask is whether the nudge results from justifiable goals. Selfish goals are difficult to justify ethically because their self-directed component makes them contradictory or irrelevant for a moral justification. Social and nudgee-driven goals however can be justified with different types of arguments that may be hold independently or in conjunction. Four types of arguments are of particular importance. First, one may argue that the nudge helps to achieve some desirable consequence³. Examples include increasing health, wealth, well-being, life expectancy, security, and political stability, or reducing suicide rates, negative externalities⁴, and discrimination against vulnerable groups. Note however that the worth or desirability of the consequences may be a matter of debate. For instance, it may be disputed whether an increase in life expectancy is a good consequence in light of the fact that aging is correlated with physical and mental disabilities (Emanuel, 2014). The difficulty then lies in showing that the intended consequences are indeed desirable and praiseworthy.

Second, one may argue that the nudge helps to fulfil important values or moral principles. Examples include respect for democratic decisions, freedom of religious beliefs, or benevolence towards vulnerable individuals. Again, difficulties may arise in cases when the values and principles put forward are controversial.

Third, one may argue that the goal stems from good intentions. Examples include motivation to care for nudgees, for vulnerable groups, or for future generations. This line of argumentation however, is

² Note that organ donation or charity giving is not necessarily against nudgees’ interest since the costs involved may be compensated in other ways. It is however against nudgees’ interest when the compensation is lower or unlikely.

³ As Bart Engelen convincingly showed in a talk given at the Designing Moral Technologies Conference (July 2016, Monte Verità, Switzerland), nudge enthusiasts use a mostly consequentialist vocabulary when they provide justifications for their nudges.

⁴ Negative externalities are damages or costs caused by a group of citizens’ behavior. For instance, citizens who do not accumulate retirement saving during their active life are likely to fall at charge of social security in their old age (for a useful discussion on this topic, see Guala & Mittone, 2015).
limited by the fact that well-intentioned motivation is not a guarantee for identifying accurately others’ interests.

Fourth, one may provide evidence that individuals affected by the nudge do or would share the goals pursued by nudgers. Examples of those evidences include results of official surveys, state voting procedures, scientific studies, and sometimes appeal to common sense. The difficulty then lies in providing convincing empirical evidence. I will elaborate on this form of argument in the second part of the paper.

The variety and quality of justifying reasons for nudgers’ goals is an important criterion for the acceptability of a nudge. In practice the quality of the justification may be assessed by investigating nudgers’ capacity to show that their goals are grounded on acceptable normative assumptions – e.g. about what consequences or ethical principles do matter – result from good intentions – rather than a patronizing and condescending posture – and would broadly find support among the individuals impacted by the nudge. Thus, the process of justifying goals is partly normatively loaded – e.g. *a priori* decisions about which consequences or principles are to be favoured – and partly empirically grounded – e.g. evidence about human basic physiological and psychological needs, or about what nudges care about.

### 2.2. Efficacy of nudges

To evaluate the appropriateness of a nudge, the second main question to ask is whether the nudge is an effective means of achieving nudgers’ goals, compared to the *status quo* or alternative ethically acceptable interventions. On occasion, in depth empirical investigation may reveal that a nudge does not produce the desired behavioural impact. The storm naming example provided in the introduction is an illustration. On other occasions, less invasive interventions, such as the mere provision of information, may have similar or better impact on individuals’ behaviour. On yet further occasions, a nudge may have too little impact compared to an ethically acceptable mandate. For instance, if the goal is to suppress smoking in public buildings, a nudge may be far less effective than an official ban.

Efficacy concerns can be alleviated with empirical evidence showing that the nudge indeed affects people’s behaviour in the desired direction. The difficulty then lies in collecting data according to scientific standards – i.e. rigorous experimental design, statistical power, etc. – and checking for their external validity (for more on this topic, see D. Halpern, 2016).

### 2.3. Ethical concerns raised by the use of nudges

To evaluate the appropriateness of a nudge, the third question to ask is whether there are concerns raised by the application of the nudge. These concerns may either refer to unwanted side-effects, or to threats against autonomous decision-making.

#### a) Concerns about side-effects

Even the most beneficial technologies may have unwanted side-effects. For instance, using defibrillators saves lives and lifesaving is an easily justifiable goal. However, many survivors of cardiac arrest suffer from severe disabilities. Thus, when evaluating the appropriateness of using defibrillators, lifesaving needs to be weighed up against the individual distress and medical needs generated by the use of this technology. Similar concerns arise with nudges. For instance, nudging women to undergo genetic testing for breast cancer may prevent many occurrences of the disease. However, it also leads to a dramatic increase of costly and invasive preventive surgery, some of which will be performed on women that would not have developed cancer.

Usually, when nudgers intervene in the interest of a specific group of individuals it is done at the expense of other individuals. For instance, a nudge designed to reduce gender biases during
recruitment procedures has positive effects for women and short-term negative effects for men’s career. Although less worrisome, such side-effects cannot be ignored.

In given social contexts, using a nudge may increase distrust against nudgers or undermine other communication efforts. For instance, a sudden change from an opt-in to an opt-out organ donation policy may raise public controversies, which in turn, may generate distrust against state officials and organ donation in general. Efficacy concerns are then raised. Similar backfiring effects may occur with nudges in favour of morally praiseworthy behaviours, such as volunteer work or charity donation: the fact of being nudged towards such behaviour may diminish its worthiness in public opinion.

Adverse side-effects matter although they may not be sufficient reasons to give up a nudge. Nudgers may recognize the side-effects while providing convincing counterbalancing reasons. In practice, weighting worrisome and positive elements is a difficult task because nudgers are tempted to overstate the positive side of the balance. However, addressing these two aspects explicitly and overtly contributes to an objective assessment of the nudge.

b) Autonomy concerns

The last fundamental question to be addressed is related to whether “nudging” is an acceptable technique to use on people. To some extent, most nudges could be said to trick people into certain decisions and lifestyles because they bypass conscious deliberation of those influenced. Since contemporary ethics places enormous emphasis on autonomous choice, the simple fact that nudgers “cheat” nudgees by framing the context of choice in ways which nudgees may not notice or take account of is controversial (Ashcroft, 2011). This general concern can be formulated differently. One may object that nudges impose a lack of transparency by clouding the availability of options accessible at the moment of the decision, and thereby endanger informed and free choice. One may conceive of nudges as unwarranted forms of manipulation or indoctrination because they exploit people’s cognitive weaknesses. One may fear that nudges undermine agency – i.e. use of deliberative capacities and control over own choices – and thereby remove the authenticity of one’s choices or the realization of one’s personality. For the sake of simplicity, I will group all these worries under the label “autonomy concern” (for more differentiated analyses, see Bovens, 2013; Nys & Engelen, 2016; Saghai, 2013; Wilkinson, 2013).

There is no simple answer to the autonomy concern. Despite their ‘soft’ character, nudges will leave a trace of moral violation as long as there are reasons to think that some nudgees would find it unbearable to be nudged. This may happen in two types of situations. First, nudgees may disagree with nudgers’ goals. For instance, individuals with religious beliefs against organ transplantation may disagree with the goal of increasing organ donation. Second, nudgees may disagree with the mere fact of being nudged. In practice, irrespective of the nudge being considered, it is quite unlikely that all nudgees would agree to be nudged if they were asked. Thus, the relevant question is rather: how much disagreement is bearable?

In order to counterbalance or alleviate autonomy concerns, at least four lines or argumentation are available. First, one may stress the importance of nudgers’ goals. For instance, emphasizing the expected gains in life quality may be a convincing reason for discarding limited autonomy concerns raised by an anti-smoking nudge. Second, one may downplay the issue by confronting academic ideals to the non-ideal world we live in: does it make sense to be reluctant against a new nudge in light of the fact that citizens are already constantly nudged – mostly against their interests – by private companies through under-regulated advertisements and marketing campaigns? Third, one may put forward what I will label a “shared preference justification” (hereafter SPJ) by arguing that nudgers pursue goals that are shared by, or in line with the preferences of individuals impacted by the nudge. Defenders of a libertarian paternalistic view typically aim at this sort of justification when they claim that nudges contribute to making people make better choices “as judged by themselves” (Conly, 2013; Thaler & Sunstein, 2008). Fourth, and more challenging than a SPJ, one may formulate a “consent justification” (hereafter CJ) by arguing that nudgees do – or would if they were asked – agree to be nudged in this
way. A “minimal CJ” may be provided by combining a SPJ with empirical evidence indicating that, in general, people accept being softly manipulated as long as nudgers’ goals are aligned with their own objectives in life. There is in fact growing evidence in this direction (Attari et al., 2009; Reisch & Sunstein, 2016). A “full CJ” may be provided by collecting nudgees’ point of view under full information about nudgers’ goals and about how the nudge works and is applied.

Shared preference justifications (SPJs) and consent justifications (CJs) are similar but should not be confused. They are grounded on different empirical data. In the former argument one is interested in individuals’ preferences related to nudgers’ goals – e.g. decrease the suicide rate in a population, make people donate their organs, or quit smoking. In the latter argument, one is interested in individuals’ evaluation of the nudge itself, that is, in the means of achieving nudgers’ goals – e.g. use of an opt-out policy in order to induce organ donation. Individuals are unlikely to consent to nudges that make them behave against their own preferences. Thus, an SPJ seems to be a precondition for a CJ. There may be exceptional situations however where this rule does not apply: if an individual is blindly convinced that a nudger acts for her good, she may give her consent even to nudges that make her behave in contradiction with her preferences.

In practice, it is easier to achieve SPJs than CJs. Empirical evidence of nudgees’ consent to a nudge may be the most convincing way to alleviate autonomy concerns, but in practice, it is difficult to gather. One would have to organize large-scale consultations about the acceptability of given policies. SPJs may require less specific information. For instance it is quite uncontroversial that people prefer a lower to a higher risk of car accident. SPJs may rely on already available information. For instance, in states where citizens can express their views via democratic voting procedures, policy-makers may already know what citizens expect from them – e.g. develop practical solutions in favour of safer roads, or renewable energy consumption. Therefore, SPJs are more likely to be proposed than CJs.

3. Shared preference justifications

SPJs may serve as justifying reasons for nudgers’ goals. Showing that all parties impacted by a nudge would mostly agree with the goals pursued by the nudgers is a powerful form of justification. In a practical ethics context, a SPJ may even play the useful role of justification-stopper. Indeed, ethical reasoning is often challenged by the regression of justification difficulty: asserting that X is justified because it fulfils an ethical principle or has good consequences leads to the further question of why this principle or these consequences themselves are justified. Reliable evidence about factual agreement on X is a form of justification that is usually understood and accepted with little controversy.

Moreover, SPJs help to take autonomy concerns seriously. Until proper investigation has been made of nudgees’ preferences, autonomy concerns remain on the table and the nudge raises doubts of moral violation and paternalistic abuse. Thus it is important that nudgers at least try to provide a SPJ. If they do not succeed, they are committed to find other overwhelming reasons to counterbalance the neglect of nudgees’ preferences.

While crucial, it is a real challenge to make a SPJ convincing because ethically relevant preferences are difficult to define and identify. Individual “preferences” may refer to different phenomena such as action-motivating states – e.g. drives, wants, actual choices – or more intellectual states – e.g. simple ‘likes’, or more complex comparative evaluations. Moreover, preferences may be about different types of states – e.g. long-term versus short-term states, global versus specific states. Deciding which of these phenomena are relevant for assessing nudges is a matter of debate.

Moreover, identifying nudgees’ actual subjective preferences – whether they are conceived as action-motivating or evaluative states – bears heavily on empirical data which are not easy to obtain because these preferences are “opaque”, “variable”, “inconsistent”, and “modifiable”. They are “opaque” in the sense that different individuals may not take the same attributes of a situation as relevant factors on which to base their preferences, and from a nudger’s perspective, it is difficult to identify the attributes
that are relevant for nudgees. For instance, a group of individuals may express a preference for giving money to a charity, some of them because they are affected by the needy state of the charity-recipients, others because they are reputation-driven and affected by the presence of third-party observers. This granularity of individuals’ preferences is not easy to investigate empirically. Preferences are “variable” in the sense that, within a population, not all individuals share the same preferences. For instance, when asked to contribute to a risky project, more risk-averse individuals may prefer to decline whereas less risk-averse individuals may accept. Preferences are “inconsistent” – at least at first glance – in the sense that the same individual may simultaneously hold preferences that call for contradictory action choices. For instance, one may both have a short-term preference for relaxing and a long-term preference for doing exercise. Preferences are “modifiable” in the sense that individuals may change their preferences over time. For instance, one may have a preference for not having children at age 20, and the reverse preference at age 35 (for more on these difficulties, see Guala & Mittone, 2015).

Furthermore, it is difficult to evaluate the authenticity of individuals’ preferences. Preferences are not formed in isolation. They are shaped by education and culture, which makes it difficult to decide to what extent they reflect individuals’ authentic preferences. For instance, in many cultures, young women have a strong preference for appearing beautiful at the expense of their time and savings (for more on this topic, see MacKenzie & Stoljar, 2000). Knowing the social pressures that lead to the formation of such preferences, how much ethical weight – i.e. relevance in a SPJ – are we supposed to give to them?

For all these reasons, it is a real challenge to identify nudgees’ relevant and actual subjective preferences in order to provide a convincing SPJ. The difficulties are particularly acute when a nudge targets heterogeneous populations which may have different goals or different understandings of their own interests. Henceforth some authors suggest simply giving up the use of such arguments (Guala & Mittone, 2015). In what follows, however, I will take on the challenge and propose a conceptual framework which will serve as basis for both the production of convincing SPJs, and the evaluation of their quality. For this, I will first (section 3.1) describe two categories of states that may be described as individuals’ subjective preferences, second (section 3.2.) I will argue for a restriction of the class of preferences that are relevant for nudge assessment, and third (section 3.3.) I will detail the three axes on which the strength of a SPJ can be evaluated.

3.1. A description of individuals’ subjective preferences

SPJ are about showing that the subjective preferences of individuals impacted by the nudge are aligned with nudgers’ goals. Preferences have been defined in multiple ways in the philosophical and social science literature (Hausman, 2011). Thus, we need to make explicit statements about what categories of preferences may be meaningfully integrated in the argument and explain the reasons for these choices.

My first choice is to focus on categories of preferences that can be held by ordinary existing human beings, as opposed to ideally rational beings. This choice is justified by the fact that SPJs involve empirical assertions about individuals actually impacted by the nudge. Therefore, in this section, I will take an empirical point of view and describe the two main types of psychological moments that ordinary nudgees do experience and that may involve “subjective preferences”: the moment of the choice between available options (Figure 1, green zone), and the moment of cool evaluation (Figure 1, red zone). In section 3.2, I will further restrict the domain of preferences relevant for nudge assessment.

---

5 I thank Chloë FitzGerald for pointing this out.
Submitted version: at this stage, any critical feedback is welcome!

Figure 1. Illustration of nudger’s versus nudgee’s point of view on a nudge.

Yellow zone: A nudger replaces ordinary stairways by luminous stairs with the aim to get people to do more physical exercise. The space of choice options remains the same in the original (A) and in the nudge (B) situations.

Green zone: Components underlying nudgee’s choice, in situation A (left) versus B (right). Nudgee’s “local motivating preferences” (i.e. ‘drives’) coupled with her awareness of the “attributes of her choice options” induce a “global motivating preference” (i.e. actual choice) for climbing stairs or taking the elevator.

Red zone: “Local theoretical preferences” (i.e. likes) involved in nudgee’s cool evaluation of the choice situations and of the nudge. Balanced against each other, they induce a “global theoretical preference” (i.e. comparative evaluation) for being placed in choice situation A or B.

“In action”, that is, at the moment of the choice between two options, nudgees recognise a number of relevant attributes of their available choice options. This evaluation may vary from one individual to the other. In the example illustrated in Figure 1, while deciding whether to climb stairs or take the elevator, a nudgee, Fatima, recognizes more or less consciously that climbing stairs involves physical effort, is good for her health and may provide her a reputation of fit person. In the nuded situation, Fatima additionally recognizes that climbing luminous stairs is a playing activity. Towards each of these attributes, at the moment of the decision, she feels – more or less consciously – a favourable or negative drive: she feels more or less motivated to have the attribute realized. Let us call these attribute-related drives “local motivating preferences”. Put together and balanced against each other, these local motivating preferences induce a “global motivating preference” which corresponds to her actual choice of action. In the original situation (A) Fatima’s global preference may be to take the escalator. While being nudged however (situation B), the additional local motivating preference for experiencing a playful activity may shift her global preference in favour of climbing stairs.

“In theory”, that is, when individuals can think over the choice situation without being urged to choose, they may recognize different attributes of the choice options and value them differently. For instance, regarding choice situation A, Fatima may disregard the reputation benefit and give a particular weight to the health benefit. Let us call these in theory evaluations or commitments in favour or against the realisation of an attribute “local theoretical preference.” After balancing her local preferences, Fatima may form a “general theoretical preference” for climbing stairs even if, in action, she would choose the escalator. In this configuration, Fatima may also have an in theory second-order...
preference that she would have corresponding motivating preferences when placed in the actual choice situation. Note that this analysis resolves some worries about individuals’ inconsistent preferences. Individuals may hold theoretical preferences that contradict their motivating preferences because, in action and in theory, they may not give the same weight to the realization of the same attributes.

Moreover, in theory, when additionally informed of the nudge policy and about nudgers’ goals, nudges may also form theoretical preferences over attributes related to “the fact of being nudged”. Thus, while evaluating the acceptability of a nudge, actors may take into account some attributes related to the choice situation, and other attributes specifically related to the nudge. For instance, Fatima may recognize that the nudge helps her having the ‘right’ motivating preference to climb stairs and evaluate positively this attribute of the nudge. But at the same time, she may have autonomy concerns increased by her lack of confidence in the nudger. Compared and balanced against each other, these theoretical preferences will induce a global preference for accepting or opposing the nudge. This example illustrates the possibility of having a theoretical preference for the nudged choice situation (Figure 1, situation B), while having a preference against being softly manipulated by the nudge.

Note that at this stage, my account involves no claim about the rationality of individuals nor about some universal components of individuals’ preferences. In action as well as in the in theory, some individuals may overlook an attribute of a choice situation that others would consider highly important. These situations can occur because the recognition and evaluation of a situation’s attributes may be influenced by environmental circumstances – e.g. stress factors –, by knowledge and past experiences, by social pressures, by cognitive biases – including those that are manipulated by nudges – by rationality impairments, or by character traits.

3.2. Preferences that do matter for a shared preference justification

In this section, I propose three features that preferences should fulfil in order to build convincing SPJs: put shortly, preferences need to be (a) theoretical, (b) over long term states, and (c) minimally rational. I will argue that this class of preferences is satisfactorily stable and representative of nudges’ interests.

a) Theoretical preferences

So far, I have provided a descriptive account of two categories of evaluative preferences. Motivating preferences participate in a more or less conscious evaluation of a choice situation and lead to actual choice. Theoretical preferences participate to the production of a conscious and explicitly stated evaluation about what the actor should – or should not – choose, or about the acceptability of a nudge.

Motivating and theoretical preferences are both evaluative in some sense, but there are several reasons to think that the latter are more directly relevant for nudge assessment than the former. First, theoretical preferences are more likely valued by nudges themselves than motivating preferences. Second, scholars working on ethical issues usually think that reflective and consciously hold evaluations deserve more weight than unconscious ones (Spellecy, 2003), and conscious evaluations are more likely to be produced in theory than in action. Third, contrarily to motivating preferences which are strictly bound to in action choice situations, theoretical preferences may be used to assess nudges directly. For instance, autonomy concerns are based on scholars’ theoretical preferences for not being manipulated. Fourth, focusing on theoretical preferences is a way to address the opacity and inconsistency difficulties mentioned in section 2.3. Since in theory evaluations involve more conscious activity than in action evaluations, it is easier to identify which attributes they take as relevant choice factors; there is less opacity among theoretical preferences. Moreover, since most cases of preference inconsistencies are situated at the interface between motivating and theoretical preferences, focusing strictly on one level resolves part of the inconsistency difficulty.

b) Preferences over long-term states
Among theoretical preferences, one may distinguish between those that are “over long-term states” and those that are “over short-term states”. The former include desires and commitments about distant future states – e.g. losing weight – or about present states that should last for a long time – e.g. maintaining a good health, or quit smoking. The latter include commitments about states in the near future, such as saving physical effort right now, or having a cigarette. Note that this distinction does not necessarily map directly onto fast and automatic versus slow and deliberative cognitive operations. For instance, a preference for not dying – i.e. over a long term state – may be produced by fast and automatic processes while a preference for climbing stairs right now – i.e. over a short term state – may partly result from a conscious health-driven calculus.

On occasions, preferences over long-term states may be fugitive or may change during an individual’s life. For instance, an over-weight child may, after being the prey to her playmate’s mockery, form a preference for durable weight loss, but after several days of not being mocked she may ‘forget’ this preference. Or a nun who, for some reason, has lost her faith may radically change her long-term lifestyle preferences.

Preferences over long-term states may be used for the formation of global preferences over short-term states. For instance, when an individual who bases her decision on preferences over long-term states is asked in theory whether she wants to save money today, she will ask herself whether she prefers enjoying life during her youth with the risk of suffering from lack of savings in her old age, or whether she prefers spending a less joyful youth in order to secure more comfort at old age.

The global preferences that individuals form in a reflective thinking over long-term states are the most relevant to take into consideration in a SPJ. Indeed, preferences over long-term states extend over actual choice situations and help to alleviate the modifiability difficulty described in section 2.3. Despite the fact that these preferences may change on occasion, they still are less modifiable over time than preferences over short-time states. Therefore, preferences over long-term states are more reliable evidence of what really matters to individuals.

3.3. How to make a shared preference justification convincing

In the previous section, I have argued that the most convincing SPF$s$ involve the investigation of preferences that are theoretical, over long-term states, and minimally rational. These preferences are
the least opaque, inconsistent, and modifiable although there is an incompressible part of volatility which correspondingly dims the reliability of SPJs. This is the price to pay in an imperfect world. Scholars working in practical ethics are used to those compromises.

On a case by case basis, the strength of SPJs may be assessed along the three following axes: (a) the degree to which nudgers’ goals align with the preferences of individuals impacted by the nudge, (b) the proportion of individuals whose preferences align with nudgers’ goals, and (c) the empirical validity of the data.

a) **Degree of alignment**

The first assessment axis is the degree of alignment between individuals’ minimally rational preferences and nudgers’ goals. A nudger may pursue various goals. Some of them are “specific”, that is, tightly linked to the expected effects of the nudge – e.g. increase statin consumption in a population – while other goals may be “broader” – e.g. decrease the rate of cardiovascular diseases in a population. Obviously, SPJs focusing on preference alignment with specific goals are more convincing than SPJs focusing on broader goals. Moreover, SPJs that provide evidence of preference alignment with goals deemed as important are more convincing.

Empirical investigation may reveal that nudgees have preferences that are “fully aligned”, “partly aligned” or “misaligned” with nudgers’ goals. For instance, a nudgee may prefer a decreased risk of cardio-vascular diseases over a stable risk – i.e. full alignment with nudgers’ broader goal – while having a preference against increasing statin drugs consumption – i.e. misalignment with nudgers’ specific goal. Or a nudgee may only weakly prefer an increase of organ donation at the national level because she does not feel concerned – i.e. partial alignment with nudgers’ goal.

Interestingly, in contemporary Western individualistic societies, nudgees are more likely to have goals related to themselves – e.g. decreasing their own risk of cardio-vascular disease – whereas nudgers are expected to have goals that serve collective interests – e.g. decreasing the mean rate of cardio-vascular diseases at the national level. One may therefore expect that “moral nudges” – i.e. those designed for improving moral decision-making such as a sustainable way of life, charity giving, or tax paying – are less likely to fare well on this assessment axis than “non-moral nudges” – i.e. those designed for improving nudgees’ health or savings (as shown empirically in Hagman, Andersson, Västfjäll, & Tinghög, 2015).

b) **Proportion of alignment**

The second assessment axis on which SPJs can fare more or less well, is the proportion of individuals impacted by the nudge that hold minimally rational preferences aligned with nudgers’ goals. As is the case with any technology, the use of a nudge impacts many individual lives. Impacted individuals include nudgees and all other individuals indirectly affected by the enforcement of the nudge and by the behavioural change of the nudgees. For simplicity, I mainly consider nudgees’ point of view in this paper, but it should be clear that a full-fledged SPJ would have to take into account the whole population of individuals impacted by the nudge.

This evaluation axis helps to address the problem of preference variability mentioned in section 2.3. Even though, within a population, individuals hold a variety of preferences, preferences that are only shared by a small proportion of nudgees may be relativized accordingly. Conversely, large proportions of nudgees whose preferences misalign with nudgers’ goals are indicators of serious autonomy concerns.

Interestingly, this assessment axis also helps to give appropriate weight to autonomy concerns. Imagine that one tries to extend a SPJ into a CT: the aim is to investigate whether fully informed nudgees would consent to be nudged. Among nudgees, some individuals may strongly dislike being nudged. They may dislike it even more than the risk of making a choice in action that they do not
endorse *in theory*. Other individuals may find it useful to be nudged because they have a predominant local theoretical preference for avoiding risk factors. Since a nudge applies to all these individuals, there is no *a priori* reason to prioritize the anti-nudge preferences. Rather, individual preferences matter to the extent that they are shared in the nudgee population (see also Nys & Engelen, 2016). Thus, it becomes interesting to evaluate the expected proportion of nudgees that have strong local theoretical preferences against being nudged *per se*. Empirical evidence indicate that this proportion is small (Attari et al., 2009; Petrescu, Hollands, Couturier, Ng, & Marteau, 2016).

Finally, this assessment axis highlights the fact that SPJs are context sensitive. The acceptability of a nudge depends very much on the overall preference structure of the targeted population. Nudgees’ preferences are impacted by all sorts of factors, including socio-cultural background or historical events. For instance, there is an important variability among countries in the way citizens see responsibility for organ donation (Jansen, Haase-Kromwijk, Van Leiden, Weimar, & Hoitsma, 2009); it follows that an opt-out policy may be supported by a SPJ in one country but not in another. Moreover, a population’s point of view may change over time. For instance, worldwide people’s perception of the safety of nuclear power has changed dramatically after the Fukushima nuclear disaster. Thus nudges designed to favour the consumption of wind or solar power over nuclear power were less aligned with citizens’ preferences before than after the Fukushima event.

c) *Scientific plausibility*

The third assessment axis for SPJs is the robustness of the empirical assertions they contain. When empirical assumptions are highly plausible – e.g. people prefer to stay in good health – there is no need for effortful investigation. When assumptions are more questionable – e.g. people are willing to donate their organs after death – there is a requirement for empirical investigation. Public consultation by means of surveys, deliberative polls, or democratic voting procedures may be required. The quality of these empirical investigations will be assessed according to usual scientific criteria: evidence-based hypotheses, testability, coherence with other established knowledge, etc.

**The limits of shared preference justifications**

Convincing SPJs aren’t knock down arguments in favour of nudges. The relevance of a SPJ depends on how much importance one gives to alternative arguments – e.g. justifications for nudgers’ goals – and how much confidence one has in the authenticity of nudgees’ preferences. At least on some occasions, one may have good reasons to think that nudgees are collectively impaired in their preferences.

Moreover, one should keep in mind that SPJs are weaker than CJ. SPJ are not a guarantee that nudgees would consent to be nudged. For instance, if a nudger is not recognized as trustworthy, nudgees may have preferences that are fully aligned with nudgers’ goals while having a preference for not being nudged by this untrustworthy nudger (Tannenbaum, Fox, & Rogers, 2014).

4. Conclusion

The aim of this paper was to propose a general framework for nudge assessment which serves as gate keeper against ethically problematic or ineffective nudges. This framework helps addressing the relevant ethical questions at early stage of reforming procedures and to balance pro and contra reasons to apply a nudge. The second part of the paper was devoted to one form of argument: shared preference justifications. SPJs are important because they may be used for three different purposes: for justifying nudgers’ goals, for alleviating autonomy concerns, and for detecting paternalistic abuses. SPJ also deserve particular attention because it is difficult to make them convincing. I have explained where the difficulties lie and proposed a way to resolve them.

---

6 I will not elaborate here on the conditions required for rigorous empirical investigation, but it is a critical and difficult step that should not be taken lightly.
Acknowledgements

Many thanks to Emma Tieffenbach, François Hudon, Chloë FitzGerald, Samia Hurst, and Bart Engelen, for their useful comments.

References


Bovens, L. (2013). Why couldn’t I be nudged to dislike a Big Mac? Journal of Medical Ethics, 39(8), 495–496. https://doi.org/10.1136/medethics-2012-101110


