Working paper
Not all nudges are automatic: freedom of choice and informative nudges

Júlia de Quintana Medina  
juliadequintana@uab.cat
Universitat Autònoma de Barcelona

Paula Hermida Justo
paula.hermida@uab.cat
Universitat Autònoma de Barcelona

Presented to the European Consortium for Political Research
Joint Session of Workshops – 2016
Behavioural Change and Public Policy
Holger Strassheim and Silke Becke Chairs
Pisa, Italy
April 24-28, 2016
Introduction

Since the publication of *Nudge* (2008) by Thaler & Sunstein there has been a global rise of behavioural informed policies and increasing attention towards nudge strategy both within academia and in public policy institutions. The growing number of nudge interventions in several public policy domains as well as the constant normative and empirical contributions from academia position nudging as a hot topic in discussions. Nudging is a type of approach to policy design that is grounded on insights form behavioural economics and cognitive psychology, which seeks to alter the decision making environment (choice architecture) and shape individual behaviour in order to improve individual and social wellbeing, while respecting freedom of choice (Thaler & Sunstein, 2008). Although this approach is gaining supporters it is also raising concerns since nudges use as public policy tools still faces legitimation issues on both its ends and its means (Mills, 2015: 496). In this paper we explore the main objections to nudging, considering both the nudge objective (the ends) and how nudges work (the means).

In the first section we focus on the objections regarding nudge political legitimation. Here, we discuss whether nudges are policy tools which achieve the same policy outcomes that other behavioural change strategies chase. In the second section, we address the objections around nudges *means* which usually highlight nudging as a manipulation base strategy that undermines autonomy, lacks transparency, and is useless for providing long term behavioural change (Conly, 2013; Gigerenzer, 2015; Hansen & Jespersen, 2013; Hausman & Welch, 2010; Quigley, 2014; White, 2013). In that respect, we suggest that the majority of ethical objections around nudges *means* are underpinned by three main assumptions: (1) an idealistic view of autonomy, (2) a focus on freedom as a normative criterion; and (3) a tendency to argue that all nudges are automatic, when in fact they differ in many aspects. We take a critical look at these assumptions to conclude that what is still lacking from the literature on nudges is a framework to address their potential for inducing persistent behavioural change. In the third section, we propose two factors that can be linked to this goal: the provision of feedback and the detachment of nudges form the execution of choice; we call this type: *informative nudges*. Finally, we present some conclusions.
A review of nudge objections regarding ends and means

Nudging is a strategy originally grounded in libertarian paternalism, an approach which seeks to change behaviour and improve wellbeing (which constitutes its paternalistic side). It does so without limiting the original set of choices, hence respecting freedom of choice (which constitutes its libertarian side). There are two normative standards most commonly employed to assess the legitimation of nudges: (1) improvement in individual wellbeing “as judged by themselves” and (2) the preservation of freedom of choice (Blumenthal-Barby, 2013; Mills, 2015). Both issues have received a great deal of attention from critics and face different objections regarding nudges and their viability as tools to change behaviour. In the following sections we revise both claims and present some counterarguments to defend the nudge approach.

1. Improvement on individual wellbeing “as judged by themselves”

The key idea of Thaler & Sunstein (2008:5) is that “choice architectures can make major improvements to the lives of others by designing environments which promote choices in individuals’ best interest”. They claim that their approach has a paternalistic aspect since “it tries to influence choices in a way that will make the chooser better off as judged by themselves” [original italics] (Thaler & Sunstein, 2008). Drawing on behavioural economic findings, they argue that nudges always promote individual wellbeing as they move people in directions they already agree with and would have chosen if they were not affected by cognitive biases, poor self-control and limited time, information and resources. Hence, libertarian paternalism relies on the subjective informed desire for wellbeing to legitimate the ends and goals of its interventions (Blumenthal-Barby, 2013). However, this point is challenging because for nudges to be loyal to this premise they must (1) identify people’s preferences and (2) accordingly organize the context(s) in a heterogenic way, which will adapt to diverse demands. Nevertheless, both of these claims are controversial, since defining what will make people better off “from their own point of view” is difficult and in some cases, a nudge could have opposite consequences on different people, with different backgrounds and preferences (White, 2013).
However, a strong line of criticism regarding nudging has focussed on the possibility of identifying people preferences and promoting them through interventions in choice architecture. In libertarian paternalism, the idea of identifying people’s real preferences is underpinned by behavioural economics and relies on the appreciation that it is usually due to cognitive biases and heuristics (and only because of them) that people fail to choose and decide in ways that will benefit their own (and others) wellbeing. Because these cognitive flaws are systematic and predictable, policymakers are able to prevent people from failing and use nudges to help them reach their true goals (Thaler & Sunstein, 2003; Thaler & Sunstein, 2008). However, this assumption has been highly questioned (Guala & Mittone, 2015; Hausman & Welch, 2010; White, 2013). The most prominent objection comes from White (2013) who questions the behavioural economics framework. The author recognizes that behavioural economics provides new insights on how people react and make choices in everyday situations, and has contributed noticeably to understanding human behaviour. Unlike mainstream economics, which in order to explain choices assumes that people have fixed and stable preferences, behavioural economics argues that preferences are instable and change according to the context. Nonetheless, White (2013) points out that despite this disagreement, both approaches rely on the preference-constraints framework, which overlooks other factors involved in decision-making for instance, emotions (Elster, 2010) and desires, beliefs, and opportunities (Hedstrom, 2006). Following this argument, it is claimed that “nudges cannot do what they promise” since policymakers are unable to identify people true interests (White, 2013:62). Hence, nudges may work to promote choices that policymakers want people to make, but fail to promote true individual decisions (White, 2013:79). Moreover, libertarian paternalism is quick to confuse decisions affected by cognitive biases from poor choices with full intention and knowledge (Mitchell, 2005). This is why, for White (2013), the main problem with libertarian paternalism is that it executes a value substitution which replaces the decision-maker’s interest with the policymaker's interest.

Regarding the potential of nudges to improve individuals wellbeing “as judged by themselves”, White’s critique sets out two different problems. The first is value substitution which concerns libertarian paternalism. The second concerns behavioural economics and its limitations in explaining behaviour. Nevertheless, it is less clear how both affect nudging as a public policy strategy to change behaviour. Regarding value
substitution, it is clear that it establishes practical problems with libertarian paternalism. However, this is not a criticism that directly affects nudging. In the literature both, libertarian paternalism and nudging have been widely used as equal terms but “while nudging is a means to promote behavioural change, libertarian paternalism is a guide or a series of constraints on what ends may be promoted” (Hansen & Jespersen, 2013: 12). Since nudges are mechanisms of intervention aimed at changing behaviour they may promote wellbeing, hence fitting in with the libertarian paternalism agenda, but they may also be grounded on other normative philosophy frameworks, and promote different ends (Hansen & Jespersen, 2013; Kelly, 2013). We find this distinction crucial to address practical and ethical concerns regarding nudges.

In terms of behavioural economics, White (2013) suggests that it is a limited framework for tackling the decision-making process since it focuses too much on preferences and overrides other sources of motivation for actions. Although we share this criticism in part, because of length and complexity, we will not analyse this point in this paper. In spite of that, as empirical experimentation demonstrates in many cases choices are indeed the result of cognitive biases and heuristics. Although the behavioural economics framework is still “under construction”, public policy can already take advantage of its insights. Firstly, since public policy today relies almost exclusively on mainstream economics and, consequently, all the interventions are usually underpinned by the idea of incentives or designed to tackle traditional market failures, using the behavioural economics framework will improve these tools, leading to more comprehensive and suitable policies for addressing behavioural change. Secondly, sometimes choices are affected by cognitive biases and heuristics and we are aware that some decisional environments (choice architecture) make us more vulnerable to their effect; we should correct those contexts to endorse decisions that are free from cognitive bias. In short, public policy should take advantage of the behavioural economics framework and work with it, improving and completing what it is already known. What is clear is that ignoring its contributions will improve neither insight on human decision-making, nor the public policy tools for changing behaviour.

Accordingly, what nudges cannot do is improve individual wellbeing from people’s own standards. Even so, as we have seen, this idea comes from libertarian paternalism and it does not mean that nudges are not suitable tools for public policy
which can be used grounded in different moral values, such as Utilitarianism and Rawlsianism (Kelly, 2013) to promote different ends (Guala & Mittone, 2015; Nagatsu, 2015). As Guala & Mittone (2015) suggest, because nudges cannot (always) promote individual wellbeing “as judged by themselves”, two options are raised for nudge defenders. The first option, is to “take a genuinely paternalistic stance and argue that nudged individuals are always better off independently of their preferences” (Guala & Mittone, 2015: 386). Although libertarian paternalism claims to focus on and improve individual wellbeing, most of the examples illustrated by Thaler & Sunstein (2008) seek to help people get healthier and wealthier assuming that is what they want (Blumenthal-Barby, 2013), hence a true paternalistic justification would be suitable. This approach raises many ethical concerns about freedom. Furthermore, the majority of authors supporting coercive paternalism disapprove of nudging that allows choice and, for that reason, is ineffective (Conly, 2013; McCrudden & King, 1999). The second option is to justify nudges from a political point of view, which means defending government interventions on individual choice in areas in which individual behaviour causes externalities and public policy problems; in other words, nudging will embrace Mills “harm to others principle” (Guala & Mittone, 2015). In these cases, the main goal of nudges is to improve overall wellbeing beyond the nudgees and correct or improve the performance of traditional regulatory tools which, for various reasons, are unable to resolve current public policy problems arising from individual behaviour. Considering this idea, nudges could be applied to two main types of behaviour that induce aggregate problems:

Firstly, bounded choices; this is individual choices affected by cognitive biases that primarily have a negative effect on individuals who engage with them. However, when a particular behaviour is repeated over a long time and by a considerable number of people it could grow into a public policy problem. This is the case, for instance of the negative individual and social consequences of consuming alcohol, tobacco and sugar. Nudges aimed to tackle these sorts of decisions have to identify the behavioural problem underpinning them to address them correctly. Moreover, it could be interesting to consider the nature of the decisions and to identify three main factors: first, how many times people encounter these decisions on an everyday basis; second, whether some behaviours have long term negative consequences that make difficult for the individuals to learn from their mistakes. Finally, if choices are compound in nature,
the more times people engage with them the more likely they are to develop negative consequences. Both the practical and ethical objections of nudging may differ considering the nature of individual choices to be addressed by interventions.

Secondly, rational behaviour that could cause aggregate problems. According to the traditional economic framework concerning public goods (characterized by non-rivalry and non-excludability) it is rational for self-interested individuals to act as ‘free riders’. Drawing on these assumptions, interventions aimed to correct this behaviour rely exclusively on incentives, bans and the privatization of public goods (Guala & Mittone, 2015). In contrast, as behavioural economics evidences, there could be biased motivations underlying what economics classifies as ‘free rider’ behaviour. For instance, in the case of recycling, while economists may suggest that self-interest leads people to avoid recycling and benefit from others’ contributions to protect the environment, behavioural economists highlight the idea that people do want to recycle but they fail to do so due to inertia, self-control problems, confusion, or myopia, etc. Along the same lines, broad evidence in social science suggests that rather than free riders individuals are in fact conditional co-operators, thus they will cooperate if others to do so and we know about social behaviour (Elster, 2010) in these cases a nudge can encourage the voluntary provision of public goods. This is what Thaler & Sunstein (2008) call social nudges and they usually mobilize and consolidate social norms (Nagatsu, 2015).

If in fact cognitive biases motivate some choices that have a negative social impact and hence creating externalities; nudges, working complementarily with other tools such as taxes, sanctions and information, can shape individual behaviour to match pro-social objectives and induce motivational behavioural change.

White (2013) argues that, nudging is not about helping people make better choices it’s about getting people to make the choices “policymakers want them to make” (White, 2013:82). Hence nudges are designed to work on individuals (Collins, 2013). We agree with this notion; however we do not see it as a problem, especially if nudges are underpinned by normative and ethically consistent principles. Moreover, the problem is not whether it is policymakers or people’s choices, but whether the choices produce ethically justified results. It is clear by now that public policy often faces
important motivational obstacles (Peréz, 2013). If we assume that resolving current public policy problems requires a change in people’s behaviour, and that traditional government tools fail to tackle them and motivate pro-social behaviours, we have to enquire into new forms and instruments that can possibly lead to better results. We are convinced nudging is one of them.

The previous considerations set out the idea that nudges do not belong to libertarian paternalism. As public policy tools to change behaviour, nudges can be used to promote much more than individual wellbeing. Under political justification and tackling the types of decisions described, nudges appear as new and useful policy tools for exploring and being applied to different public policy domains. In short, nudge ends are not pre-established, and may be to promote the goals and outcomes that governments seek to achieve through other behavioural strategies such as information disclosure, subsidies or taxes. Nevertheless, it is argued that a number of ethical issues regarding nudges are overlooked if they are seen as “just another mode of influence in the toolbox of state controls” (Baldwin, 2014:832). Hence nudge means, still need to be reviewed and we will tackle this question in the following section.
2. The preservation of freedom of choice: three questionable assumptions

The second focal point of the Thaler & Sunstein approach is their libertarian character. The authors argue that, in general, people should be free to choose what they want to do according to their preferences. For that, libertarian paternalism “attempts to design policies that increase or maintain freedom of choice” (Thaler & Sunstein, 2008:5). Hence libertarian paternalism measures, including all types of nudges, should be liberty-preserving “want to make it is easy for people to go their own way; they do not want to burden those who want to exercise their freedom” (Thaler & Sunstein, 2008: 5). The definition excludes nudging from being a coercive strategy, since it does not change economic incentives, does not ban or exclude any options and enables those who want to opt out to do so (Mills, 2013; Quigley, 2013; Sunstein, 2015). Nevertheless, several authors argue that Thaler & Sunstein’s idea of liberty, as an opposite of coercion, is narrowed to capture the possible ethical implications that nudges might have on behaviour. In that respect, the majority of critics agree that just because nudging offers a degree of choice this does not mean that it is respectful of individuals’ liberty and decision-making capacities (Goodwin, 2012; Hansen & Jespersen, 2013; Hausman & Welch, 2010; Kelly, 2013). In terms of the debate, what seems to be the real concern about nudging is how it exercises its influence. “The crucial point is that even if choice architecture does not block or significantly burden choices it might still interfere with a person’s ability to discern and consider options and act according to her own preferences so it might interfere with her autonomy […] Moreover, if nudges do not engage rational argument and instead bypass reasoning by exploiting non rational elements of psychological processing and or by influencing choices in a way that is not obvious it constitutes an act of manipulation that also undermines individual autonomy” (Blumenthal-Barby 2013:190) [our bold]. As this quotes encapsulates, whether nudges imply manipulation and how they affect autonomy are both major ethical concerns for discussion (Goodwin, 2012; Hansen & Jespersen, 2013; Hausman & Welch, 2010; Kelly, 2013). Objectors have persistently claimed that nudges work by manipulating people’s choices (Hansen & Jespersen, 2013; Wilkinson, 2013) and as a result they represent a threat to autonomy. To this extent, it is common in the field to argue that nudging always works bypassing people’s capacity for deliberation and in order to be effective this process must be hidden from individuals since “nudging typically works best in the dark” (Bovens, 2009: 3). For both reasons,
individuals have lost control over the evaluation of options and the deliberation of choice, hence personal autonomy is undermined (Hausman & Welch, 2010). The main critics (e.g. Bovens, 2009; Hausman & Welch, 2010; White, 2013) suggest that both manipulation and covert influence are the main intrinsic characteristics of nudges. More specifically it is argued that nudges work through covert manipulation to steer people away from their own choices and autonomy so that they then engage in behaviours which satisfy the nudger interest. To this extent, individuals no longer have reasons on which to base and explain their behaviour (Bovens, 2009). This approach is been understood as being as coercive as traditional paternalistic policies because it has an intrinsically covert nature and the effect of that means it is easier to use to promote unethical results (Conly, 2013; White, 2013). Following these arguments, many authors (Bovens, 2009; Hausman & Welch, 2010; Mitchell, 2005; White, 2013) distinguish rational persuasion from nudging, excluding any type of intervention that engages reasoning to qualify as a nudge. As we will discuss in the following section, we suggest that this assumption is wrong.

In line with these arguments, another important objection draws on the consequences of nudges, both short and long term. The key question is whether nudges lead to a decrease in individual responsibility in decision-making. Although this topic has been developed less, since it requires more time and empirical testing (Bovens 2009); nudges detractors maintain that nudges do not help us to learn to make better choices in the future because they override our conscious reasoning and block the learning process (White, 2013:102). This conception is underpinned by, as we will discuss in the following section, the mistaken idea that nudges are intrinsically manipulative, and always work automatically, meaning bypassing our reasoning capabilities. Because of that, nudging infantilizes individuals and undermines their responsibility (Bovens, 2009; Gigerenzer, 2015; White, 2013). Therefore, since nudging reinforces cognitive biases, and does not allow individuals to learn from their mistakes it is an ultimately counterproductive strategy, with regard to the capacity for decision-making, which “does not have the ability to create sustainable effects on people’s behaviour for the long term” (Bovens, 2009:11). Bovens (2009) also discusses the possibility that nudges induce long term preference changes due to cognitive habituation. After a long period of engaging in actions encouraged by nudges, individuals might naturally develop preferences for the (nudgers) preferred options.
which will lead them, for instance, to develop tastes for healthy foods and therefore, increase their consumption of them. Nonetheless, this option only contemplates the possibility of adapting tastes and preferences automatically; something we can anticipate will require a long time, since ultimately will be a result of unconscious adaptation. Moreover, this process assumes that all nudges will always work; ignoring the reality that they usually fail to induce social transformation due to the fact that they are easy to overcome (Conly, 2013; Mills, 2013). Neither scenario contemplates the fact that some types of nudges can bring awareness to our everyday choices and decisions and may encourage short term and long term behavioural change.

**Assumptions about nudge means**

We suggest the majority of ethical objections around nudge *means* are underpinned by three main assumptions. Firstly, an idealistic view of autonomy; we argue that the general sense of autonomy used by the objectors is rather idealistic since it implies that internal and external causes for actions are perfectly discernible. Accordingly, internal causes, which are assumed to be mediated and result of deliberation, should be always respected and prioritized. Secondly, the whole ethical and normative debate on nudging revolves around the idea of freedom, as it ought to be the only normative criterion to take in to consideration when addressing the normative adequacy of behavioural change interventions. Hence, while there is an extended debate on the effects of nudging on freedom of choice and autonomy, its implications on equality and adequacy are often neglected. Finally, critics address their objections considering that *all nudges are automatic* and hence they function with the same pattern of influence that is, bypassing people’s capacity for deliberation and doing so in an unrecognizable way for the nudgees. In the following section, we take a closer, critical look at these three assumptions to argue that some important considerations are missing from the fundamental arguments on whether nudging is an overall good or bad strategy.
2.1 A critical view of the concept of autonomy (Assumption I)

When addressing the ethical discussion around nudging, autonomy is the key value that is assumed to be at risk. As we have previously pointed out, nudge detractors highlight that choice architecture interventions make individuals passive agents and steer people away from their capability to act for a reason (Bovens, 2009) and to control their own decisions (Hausman and Welch, 2010), since nudges “interfere with a person’s ability to discern and consider options and act according to her own preferences” (Blumenthal-Barby 2013:190). However, these objections rely on a rather idealistic and heroic view of autonomy which assumes that individuals are always able to engage in reasonable thinking and act according their preferences, their will, or a supposedly ‘rational’ meaning resulting of deliberation. Moreover, these notions of autonomy presuppose a level of self-knowledge and internal deliberation that is difficult to grasp in real decision; since we often are not completely aware of what motivates our actions (Felsen & Reiner, 2015; Mills, 2013; Schubert, 2015). This assumption is closely linked to the idea that there are purely internal causes for action that should be respected and preserved to guide individual decisions. More concretely, we suggest that nudges detractors presuppose that internal and external causes for action are perfectly distinguishable; thus nudges are external (and intentional) influences that disturb pure and good internal reason to act. For that reason, nudges should be rejected. Nevertheless, choices are always the result of causal influence; they are never ‘free’ in the sense of being ‘non-causally influenced’. Moreover sometimes internal causes and preferences lead to systematic mistakes which have important policy implications. Our take is that, in those cases, interfering with internal causes does not matter that much.

In fact, critics on nudging who enthusiastically point out the limitations of both rational theory and behavioural economics, and strongly disagree with the ideal of Econs as utility maximizers as being an unrealistic framework for understanding human decisions, paradoxically rely on an ideal notion of autonomy which assumes a perfect conception of reasoning and decision-making, something which has been proven to be unrealistic (Felsen & Reiner, 2015; Mills, 2015; Schubert, 2015; R. Thaler & Sunstein, 2008). Relying on this super-agent understanding of autonomy (Schubert, 2015:13), the majority of nudge detractors even defend rational persuasion, and education as the ideal ways to influence and induce behavioural change (Gigerenzer, 2015; Hausman &
Welch, 2010; White, 2013), strategies that in fact fit in perfectly with the classical economy framework. Evidence shows that, in some cases, these strategies do not always work, and do not induce behavioural change in the short term (Datta & Mullainathan, 2014).

In the light of the findings on behaviour, the fact that humans suffer from bounded rationality and bounded individuality (Schubert, 2015:13), and the certainty that multiple factors (both internal and external) influence our own preferences, values and conceptions of the world, it seems naive to assume that internal causes are the only desirable way to achieve autonomous decisions. Choices, values and actions are, more often than not, not the consequence of deliberation based on purely internal motives; but the result of a combination between preferences, will, context factors, past experience, emotions, etc. Considering these findings, there is growing agreement on the assumption that real autonomy requires a combination of internal and external conditions (Mills, 2013, 2015). Therefore, the key discussion is not (only) on autonomy. We should also discuss whether we accept any external intentional influence as long as internal causal factors are not overridden. Our answer is that indeed we should. To us, well designed nudges work as external intentional sources of influences that will ensure a better relation with internal influences leading to an increase in personal autonomy. As Mills suggests, “some instances of choice architecture are not only compatible with personal autonomy, but can promote it, ‘the decision-making situation’ can be designed to protect authentic choices so that it contains options that the individual would choose in a situation free from obstacles (either internal or external)” (Mills, 2013:31).

2.2 Freedom as the only normative standard (Assumption II)

The previous consideration of autonomy is linked to the idea of freedom as the main normative standard to consider when evaluating public policy interventions to change behaviour. Nevertheless, there are no specific normative reasons why a normatively justified aim should always be achieved through people’s ‘free choices’. As Pérez (2013) argues, “we should not take individual freedom to be the decisive criterion of differentiation between permissible and impermissible motivational instruments. Instead we should evaluate motivation strategies less according to their effect on
individual freedom and more in line with the extent to which that regulatory instrument can be publicity checked and controlled” (Pérez, 2013: 3). Other normative standards also appear to be relevant when addressing the normative validity of nudging. Regarding state interventions on different domains, it has been discussed both theoretically and empirically, how they impact freedom of choice, as well as other, normative relevant justifications; to conclude that event though freedom of choice is compromised, other equally relevant normative standards are promoted. Hence, in some cases, we encounter a trade-off between different goals. For that, we suggest that more comparative research between nudges and other motivational strategies is needed to assess, not only their implications for freedom (freedom of choice and freedom of will) but also their pros and cons in terms of equality, justice and adequacy.

2.3 All nudges are automatic (Assumption III)

Recently more attention has been given to the idea that normative conclusions around nudging should consider how nudges are designed and how they work. According to the literature that explores this question (e.g. Hansen & Jespersen, 2013; Baldwin, 2014; Fischer & Lotz, 2014) the nature of the behaviour and the decisions targeted by nudges (Fischer & Lotz, 2014), the type of reaction that is expected, and whether the nudge is visible for the nudgees (Hansen & Jespersen, 2013) are some of the main factors to be considered.

One of the most important elements to discuss is whether nudges require some degree of rational engagement. There is widespread intuition that envisages a distinction between nudges that imply a conscious or unconscious reaction; thus, nudges that involve some degree of rational persuasion and nudges that work automatically. According to the Dual Process Cognitive Theory, Sunstein (2016) distinguishes between interventions depending on whether they engage System 1 or System 2. However, he considers that more research is needed to correctly define this distinction and recommends assessing the ethical objections case by case and taking in to account the nudges’ consequences. Nevertheless, following this idea, a distinction can be made between interventions which require little conscious engagement (that mainly work via automatic or non-conscious psychology processes) from those that work in a more
conscious way (Hansen & Jespersen, 2013; Hollands et al., 2013). Chiefly, interventions aimed at reducing calorie intake may require little to no conscious attention, for instance, when changing the size of a bowl to automatically reduce the calorie intake, or when healthy foods are placed in more accessible areas; whereas they might involve a more conscious response when product labels include calorie information to raise awareness of the risk of consumption in order to change food choices.

An epistemic distinction on this question has been made by Hansen & Jespersen (2013) who have elaborated a typology which distinguishes between type 1 and type 2 nudges. Whereas the aim of type 1 nudges is to change behaviour without involving reflective thinking, type 2 nudges involve reflective thinking and their main objective is to influence behaviour that is the result of some degree of deliberation. In their definition, both types of nudges are designed to appeal to both system 1 and system 2 “but while type 2 nudges are aimed at influencing the attention and premises of – and hence the behaviour anchored in – systems of reflection system by influencing the automatic system, type 1 nudges are aimed at influencing the behaviour caused by automatic thinking or consequences therefore without involving reflective thinking” (Hansen & Jespersen, 2013:14). In their typology, some examples of type 2 nudges include: the “look right” warning in the streets of London, the “fly in the urinal”, and the energy bills which use social comparison of consumption. On the other hand, some examples of type 1 nudges include, the reorganization of a consumer environment to make (un)healthy food more (in)visible or (in)convenient to find, the alteration of the size of the food recipient, and the automatic default rules that prompt people to sign up as organ donors. From our point of view, their classification encapsulates a key question: whether interventions require an active or passive reaction of the nudgees to succeed. In this sense, the degree of conscious engagement to understand the interventions may differ from nudge to nudge. For instance, food labels on unhealthy products (aimed to prompt informed food choices for healthier consumption) may use an easily interpretable format, such us the traffic light system; or may require more system 2 engagement when exposing the product’s relative calorie count in relation to the recommended daily calorie consumption. Nevertheless, both cases require the nudgee to notice the intervention and actively react to it; hence type 2 nudges always require an active response. Type 1 nudges work automatically, type 2 nudges
require reflexion. This distinction sets out an important point against criticism of them. Nudges as external and intentional causes do not always work through the same degree of influence and do not always work unconsciously; some of them require an active response from individuals. We find this topic a question that requires more discussion, both theoretically and empirically, to address the ethical implications of nudges and to describe how they work in greater detail.

A second important issue on the discussion of nudge means is whether nudges are noticeable and whether they allow the nudgee to identify both the intervention and its intentions. To refer to this idea, critics usually use the term ‘transparency’. To address this point we suggest that a distinction should be made between transparency regarding nudge ends and nudge means. Transparency of nudge ends should be understood as a public policy requirement. As public policy tools, nudges should be subjected to the same control mechanisms that other behavioural change tools go through; hence, taking in to consideration the outcomes policymakers want to pursue through them\(^1\) and accepting them through traditional democratic means. On the other hand, transparency of nudge means encapsulates whether the means and objectives of an interventions are noticeable for the nudgees.

Hansen & Jespersen (2013) also develop this question; to their first criteria (type of nudge) they add a second dimension: transparency. In their framework, a transparent nudge is as a “nudge provided in a way that the intention behind it, as well as the means by which behavioural change is pursued could reasonably be expected to be transparent to the agent being nudged” (Hansen & Jespersen, 2013: 17). Hence, it is a question of means transparency. Using this dimension they argue against the persistent objection in the nudge literature which assumes that nudges always operate in the dark and they have an intrinsically covert nature, since if they were visible to the nudgees they would most likely fail to change behaviour (Bovens, 2009; White, 2013). They successfully argue that interventions on choice architecture can be both noticeable and unnoticeable and this is independent of their effectivity (Hansen & Jespersen, 2013:19). Hansen &

\(^1\) Empirical studies on the social acceptability of nudges show that rather than reacting against how nudges work, to evaluate their tolerability people mainly consider their goal. For more on this see: Sunstein, 2015.; Sunstein, 2016; Tannenbaum, Fox, & Rogers, 2014)
Jespersen’s (2013) take on transparent nudges distinguishes among interventions that allow the nudgees to perceive the nudge and reconstruct its effect on personal behaviour. Both, type 1 and type 2 nudges can be transparent and non-transparent. As transparent nudges they identify the previously listed type 2 nudges (the look right warning, the fly in the urinal, and The Home Energy Report). Some other examples of type 2 transparent nudges include: calorie listing on restaurants menus, labels with calorie information, and seat belt alarms. On the other hand, some examples of non-transparent nudges include the previously listed type 1 nudges, as well as, implicit visual illusions in traffic such as painted speed bumps to reduce speed, and some examples of self-enrolment programmes for saving and retirement that do not require choice. Regarding type 2 non-transparent nudges they use as their main example framed medical treatment information.

Although Hansen & Jespersen (2013) provide a framework to classify examples of each case: type 1 transparent/non-transparent and type 2 transparent/non-transparent nudges, it appears to us that type 1 non-transparent nudges, and type 2 transparent nudges are more frequently found and more consistent in terms of what they want to achieve and how they achieve it. For instance, in their classification “announcing on time when trains arrive on time” and “playing of relaxing music while passengers board a plane in order to calm them” count as type 1 (unconscious), and transparent nudges. We believe that rather than transparent they are perceptible, indeed people can notice the music or hear the “on time” notification; however it is less clear whether they will perceive the effect those interventions have on their behaviour. For that reason, their nature as type 1 nudges is much more important for explaining the resultant behaviour. When a nudge is perceptible but the nudgees cannot recognize the influence over their own behaviour does means transparency matter? We think that to be properly means transparent both the nudges’ presence and their effect on behaviour ought to be noticeable. And the same time, if we want to motivate a conscious reaction to set of options what is the point of making an intervention non-transparent? For that reason, we question whether reflective and non-transparent nudges make sense. Seeing Hansen & Jespersen’s examples we think that these categories are rather weak. For instance, as a type 2 nudges they propose the following example. “Framed medical treatment

2 We are doubtful about whether this is in fact a nudge or a provision of factual information, since it is difficult to see the behaviour this nudge want to provoke.
information”; while the decision after receiving framed information of medical treatments requires some degree of conscious engagement, if the intervention is unnoticeable and the behaviour which it tries to pursue cannot be identified by the nudgee, then the nudge is not recognizable as an intentional external influence and the nudgee reaction will be similar to a decision resulting from a type 1 nudge.

According to these considerations, relying on the factors mentioned above (conscious engagement, active or passive response, and recognition) we suggest that it could be interesting to explore a distinction between two ideal types of nudges: automatic nudges and reflective nudges. Automatic nudges do not require conscious engagement and deliberation; they motivate a passive response, and are not easily recognizable by the agents. Reflective nudges, require a conscious engagement and involve some degree of deliberation; they motivate an active response, and both their means and ends are to some degree noticeable and recognizable by the nudgees.

Bringing back the notion of autonomy previously mentioned it is important to understand that both types of nudges will be ethically consistent in terms of autonomy. To us, automatic nudges and reflective nudges appear ethically consistent. That is, in the sense that they act as external intentional causes of influence which interact with internal causes to simplify certain outcomes, thus preventing individuals from making poor choices, and ultimately enhancing individual behaviour. Critics may argue that automatic nudges are unethical and much more coercive than traditional paternalist policies because of their intrinsically covert nature. However, one may ask the opposite question: are we sure that traditional policies such as bans, etc. are always overtly presented and understood by the individuals who have to comply with them? Do they allow individuals to understand how much they affect their behaviour? We doubt it; all the debates on behavioural effects and the informative power of law draw on this. Some may argue that they are better in terms of transparency ends (Pérez, 2013), but the difference between bans, informational campaigns and automatic nudges on transparency means is not that clear.

The second prominent problem with this distinction is that following restrictive definitions of nudging (Hausman & Welch, 2010) reflective nudges may not count as nudges. However they do fit in with the original definition because they are liberty
preserving; they do not significantly change economic incentives; they do not involve other methods of persuasion; and are designed according to findings of behavioural economics (Oliver, 2015; Thaler & Sunstein, 2008). Moreover, following Hansen’s (2016) revised definition of nudge, interventions of both types of previously defined as ideal types “significantly alter or influence the behaviour of Humans, even though they would be ignored by Econs” (Hansen, 2016: 3) will count as nudges. In this sense, a lot of interventions provide information which goes beyond neutral and factual information. For instance, through information, interventions can raise the salience of certain possible outcomes, can provide feedback that makes the hidden consequences of an action apparent, and can communicate social norms to change people’s expectations about how other people behave in similar situations and thus, induce pro-social behaviour. We suggest that information-based interventions such as energy labels (on cars, and home appliances) and nutrition labels on unhealthy products fit the nudge definition. Traditionally, these interventions have been seen as ways to address market failures (lack of information), however they can also be interpreted as preference-shaping. To this, Econs with complete information but stable preference on, fast and highly-polluting cars will indeed ignore energy labels information.

Taking in to account these considerations, what seems to be the main difference and problem between automatic nudges and reflective nudges has to do not with autonomy but with agency and the active role an individual has through the decision-making process, since “active choices are a way to constitute our identity” (Schubert, 2015). In the line of Bovens (2009) idea of infantilisation, Schubert (2015) points out that nudges may lead to excessive convenience, and rather than taking away our autonomy, they systematically discourage active choice which is critical in a person’s character formation. For that, nudges (especially automatic nudges) will undermine our ability to learn from our own preferences when we have to use them in a particular context and over time. This distention between autonomy and agency raises new implications for nudges. First, we have the possibility to identify life domains with different degrees of agency involvement. For instance, organ donation is a topic which we can expect raises moral and deep views and implications about life and individual agency and character. For that, a default (nudge) will not be the best option to facilitate a decision. On the contrary, food choices are usually unintentional and less (not) constitutive of our character, hence a nudge (both automatic and reflective) has less
involvement for individual agency. The second important point is the idea that nudges indeed may help us to learn new preference through active choices.

To this point, we have seen that nudges differ in many aspects. Until now the distinctions have focused on: the degree of conscious engagement, the awareness of the interventions, whether the type of response is active or passive, hence the involvement of agency; and finally the type of decision (intentional or unintentional). To us, it seems that two important issues are still lacking in the classification of nudges: 1) the degree of reflection over one’s behaviour 2) the dependency of nudges on choice execution.

To us, both ideas are key to understanding a focal point: the degree to which nudges allow the nudgees to understand the connection between their behaviour and its (negative personal or social) consequences. In our opinion this is a key question for addressing nudge adequacy as a strategy for behavioural change. This, to us, represents one of the main flaws regarding traditional tools for behavioural change. They do not allow the individuals an association between individual behaviour and both personal and social consequences. We suggest that, traditional coercive and more libertarian measures to change behaviour distance the individual origin of the behaviour from its results and consequences. For instance, bans exclude individual responsibility and self-control. Similarly, the main goal of taxes is punishing a choice by economic incentives but the consequences are excluded from the decision making context; and finally informational campaigns are often not personalized and usually reinforce individual cognitive biases (Datta & Mullainathan, 2014). We suggest that if nudges allow the connection between behaviour and consequences then this could indeed be a promising path to true behavioural change with both, short and long term positive consequences.

3. A hypothetical path to long term behavioural change: reflexion and independence of choice execution

How nudges induce persistent behavioural change is one of the pending topics to discuss. As we have previously argued, considerations on this topic draw on the consequences of *automatic nudges* and neglect the long term effect that *reflective nudges* have on the behaviour. In response to this, we suggest that *reflective nudges* which require active choice, some degree of conscious engagement, and enable nudgees to identify both, the intervention and whether it has an effect on their behaviour, are
nudges with the potential to make people more aware of their cognitive limitations. We suggest that nudges which engage with the rational deliberation system (system 2) are more likely to involve a conscious learning process and more likely to result in true preference change, also referred to as habit formation. For that, we emphasize the necessity to evaluate nudge policies effects. Some reflective nudges can allow individuals to realise the hidden consequences of an action, as well as the implications of individual behaviour in a certain outcome. To us, identifying this causal relationship constitutes a first requirement to induce a truly persistent behavioural change. Nevertheless, it has to be distinguished between (1) a nudge that makes people more aware of their cognitive limitations and the relation between individual behaviour and consequences, and (2) a nudge that will allow a true preference change. We suggest that the best way to allow this is through the provision of feedback.

The second requirement is on the relation between nudge, choice architecture and the behaviour they want to change. Nudges are originally designed to alter an immediate behaviour, that is to say a behaviour being experienced in a particular environment. For that reason, they are usually attached to a particular choice set. This is the general case for both, automatic nudges and reflective nudges. No matter how much active response and degree of conscious engagement is required, nudges work because they are attached to the choice set. This implies that when the nudge is removed from a particular context its effect will, most likely, not persist. This question closes the gap between practical and normative objections to nudging since it captures an important topic of discussion: what can we expect from nudging as public policy tools to change behaviour.

To address this, to the previous distinction (automatic and reflective nudges), we add a last criterion: whether an intervention is attached to a particular choice execution. That is, whether nudges ought to be present or not when a choice is taking place; or to the contrary, whether they work without the necessity of being present in the choice set. We suggest that, when the intervention on the choice setting is not mandatorily linked to choice execution it detaches the nudge effect from a particular choice, hence it can provide a reflexion on the whole behaviour and after a time, the behaviour will persist without the necessity for the nudge to be present. Take as examples reflective nudges with the same objective: to reduce energy consumption. A sticker next to the electricity
switch works requiring active behaviour, some degree of conscious engagement, and it is a transparent intervention since it is easy for the nudgee to identify the intervention, and the effect it has on their behaviour. Nevertheless, this nudge requires the sticker to be present and linked to the choice execution (switching the light off). Similarly, labels that indicate the benefits of low energy light bulbs ought to be present (on the product) when the buying choice is being executed. On the other hand, the Home Energy Report by OPOWER (Allcott, 2011) that compares household consumption in a neighbourhood using social norms is independent from a particular choice execution. Instead it has the ability to affect more than one choice execution; for instance, switching off the lights, reducing the water consumption, opening the windows instead of using the air conditioning, etc.

We selected this example because it is been proven to have a persistent long term effect (Allcott, 2011; Frey & Rogers, 2014). The OPOWER interventions lasted for 2 years in selected households and for 5 years in other households. Every month, the Home Energy Report provided: social comparison of consumption (which mobilized social norms to change behaviour), simpler information about consumption (explained in graphs) and action steps to change behaviour in order to reduce energy consumption in the following months. The results showed a significant change in household behaviours, and a persistent effect even without the intervention. The experimental groups in which the intervention stopped after 2 years maintained a similar energy consumption pattern. Frey & Rogers (2014) describe 4 “persistence pathways” to explain the persistence behavioural change in household’s energy consumption after the Home Energy Reports sent by the energy-efficiency firm OPOWER. These pathways are: habit formation, changing what and how people think, changing future costs and harnessing external reinforcement. Their article also suggests the possibility of identifying additional pathways. Although underlying the 4 pathways there are different mechanisms that explain behavioural change and perception of change, one key factor that enables reflection and success is feedback.

To us, the Home Energy Report by OPOWER is a great example of a nudge intervention that involves agency and has positive results because of that. Through feedback it involves reflection on personal behaviour and its consequences, which allows the nudged households to observe and experience the direct relation between that
behaviour and its consequences thus, between behavioural change and significant reductions in energy consumption. As Frey & Rogers (2014) brilliantly point out in their article, the pattern of persistence can work through more automatic reactions, such as the habit of switching off the lights automatically after leaving a room. It could also be motivated by reflexion, for instance, because we start to perceive ourselves and those around us as “energy savers” implying a modification of self-perception and identity or because a social norm around energy consumption emerges, and we want to follow it. However, the necessary condition is to observe how our everyday behaviour, even in the small things, impacts on overall energy consumption. But this persistence is also possible because the nudge is not attached to a choice execution, hence, people cannot rely on it as a clue to change their immediate behaviour, and instead their choices are a result of the adequate integration of the nudge. Part of the success of the Home Energy Report is that the nudge includes several other nudges, for instance: simplification, social norms mobilization, information on causes and consequences etc. This is possible because it is an intervention that can be separated from the choice execution and can therefore mobilize other references that impact on behaviour.

Drawing on these considerations, we suggest that to allow true behavioural change through nudges, it would be interesting to explore interventions that can work detached from 1) choice execution, and that through information 2) provide feedback. Both characteristics could allow nudges to have a persistent long term effect which is one of the main challenges surrounding nudges as public policy tools to change behaviour. We call this an informative nudge.

Nudges are interventions aimed to change behaviour. Nevertheless, the vast majority of experimental research in the field has aimed to envisage how different nudges perform in different contexts, and so the main focus is on immediate effects. It is less usual to examine how those effects persist; that is, whether after the nudge and without it those who had experience it continue to behave differently (Frey & Rogers, 2014). For us, if nudges result in persistent behavioural change they should be a key topic for discussion and empirical testing. We do not know (yet) if nudges will lead to the desired effect, and we do not know how this is going to translate to short term and long term choices, preferences and overall capacity to make decisions. This is a highly controversial topic since there is no empirical evidence available and gathering...
information about such factors through empirical data could prove quite a challenge. Taking these limitations into consideration, we suggest that the idea of informative nudges; that is, those that provide feedback and work independently of the choice execution; could guide future experimental research, that could help us find out the short and long term effects of interventions, and identify those interventions that allow active, conscious, and persistent behavioural change.

Conclusions

Objections around nudging have been prominent in the last 10 years and have led to a variety of conclusions on the practical and normative viability of nudges as tools to change behaviour. This paper sets out some of the limitations of these objections and points out that some important considerations are missing from the fundamental arguments on whether nudging is an overall good or bad strategy. To that end, it is suggested that many considerations need more theoretical and empirical attention.

Regarding nudge ends, the paper draws on the necessity to provide a true political justification on nudging that legitimates its use as a useful public policy tool. The discussion suggests that nudge ends are not pre-established and can be in accordance with the promotion of the goals and outcomes that governments seek to achieve through other behavioural strategies such as, information disclosure, subsidies or taxes. What seems important in this discussion is the type of decisions they should be tackling. In this line it is suggested that nudges should be used to correct individual behaviour which primarily has a negative impact at a macro level, this is that creates externalities. Moreover, some considerations around the nature of the decisions seem important for addressing some of the ethical objections. This is an area that needs more research and could be interesting to develop in the future.

Regarding nudge means, the paper sets out some important limitations of the main objections that see nudging as a manipulative and ineffective strategy in the long term. Firstly, the main objections regarding autonomy rely on an ideal view of the decision-making process and are overly concerned with the effect that nudging has on freedom (of choice and of will). Well-designed nudges work as external intentional
sources of influences that will ensure a better relation with internal influences leading to an increase in personal autonomy. Moreover, sometimes we encounter a trade-off between freedom, and equality, or adequacy of certain strategies to change behaviour. For that, rather than developing more discussion on freedom it could be interesting to explore the advantages and disadvantages of nudging with regard to other normative public policy standards. Secondly, the paper argues that not all nudges work in the same way. For that, their implication in the decision-making process; and on the agency of the individuals ought to be discussed considering how nudges are designed and how they work. Following this argument, the paper suggests two ideal types of nudges: automatic nudges and reflective nudges. In this paper, we are not able to discuss their insights and implications in any depth; hence a closer look into this question is needed. In future research we are keen to discuss how the degree of conscious engagement and the noticeability of an intervention alter and affect the overall nudge performance.

Thirdly, a gap in the nudge literature is exposed: what seems to be fundamentally lacking in the nudge discussion is a proper take and reflexion on their consequences. To address this issue, two nudges features are identified to be relevant: whether or not nudges provide feedback, and whether or not they are attached to choice execution. A reflective nudge that has both characteristics is called an informative nudge. These are two main features that, as far as we can see, have not been greatly explored in the field. To do so, we suggest that: on a theoretical level, discussion on nudge consequences should draw on reflective nudges, not on automatic nudges and rather than focussing on the effects nudges have on autonomy and learning abilities, they should involve a deeper discussion on agency and the consciousness of individual roles on the outcomes. In that sense, feedback appears to be a key factor to enable this process. On an empirical level, it would be interesting to try and test interventions that are independent from choice execution. It is argued that, when nudges are not linked to choice executions, if they succeed it is mainly because the active role of the agent who is aware of the impact that their individual behaviour has on a certain outcome. The intervention is therefore much more likely to persist, which is ultimately an important goal of public policy.
Bibliography


