Rev.Phil.Psych. (2015) 6:481–494
DOI 10.1007/s13164-015-0245-4

# Social Nudges: Their Mechanisms and Justification

**Michiru Nagatsu**

Received: 21 October 2014 / Accepted: 27 February 2015 / Published online: 25 March 2015


**Abstract** In this paper I argue that the use of *social* nudges, policy interventions to induce voluntary cooperation in social dilemma situations, can be defended against two ethical objections which I call objections from coherence and autonomy. Specifically I argue that the kind of preference change caused by social nudges is not a threat to agents' coherent preference structure, and that there is a way in which social nudges influence behavior while respecting agents' capacity to reason. I base my arguments on two mechanistic explanations of social nudges; the expectation-based and frame-based accounts. As a concrete example of social nudges I choose the "Don't Mess With Texas" anti-littering campaign and discuss in some detail how it worked.

## 1 Introduction

Thaler and Sunstein (2008) popularized *libertarian paternalism*, the idea that behavioral economics and psychology —"the emerging science of choice" (p. 7)—provides policy makers with new tools to influence people's economic and other choices for their own benefit without compromising their freedom of choice. Similar ideas have been proposed under different banners, such as 'asymmetric' (Camerer et al. 2003) and 'light' (Loewenstein and Haisley 2008) paternalism. I call these ideas collectively *nudge paternalism*. Its core idea is the use of nudges. Nudges are subtle behavioral interventions that are distinct from standard regulations that operate with incentives. Although nudges have already been applied as *behavioral public policy* in a wide range of domains (Shafir 2013), nudge paternalism has attracted ethical and moral

M. Nagatsu (✉)
TINT, Social and Moral Philosophy/Department of Political and Economic Studies,
University of Helsinki, PL 24 (Unioninkatu 40 A), Helsinki 00014, Finland
e-mail: m.nagatsu@gmail.com

criticisms (e.g. Bovens 2009, Sugden 2009, Hausman and Welch 2010, Grüne-Yanoff 2012). These critics, as well as the proponents of nudge paternalism, have not so far paid sufficient attention to (i) the different types of psychological and social processes exploited in paternalistic interventions and (ii) the implications of such differences for the normative evaluation of these interventions (but see Grüne-Yanoff 2014). Specifically, existing discussions of mechanisms of nudges tend to focus on individual psychological mechanisms rather than collective and interactive mechanisms, and on agents' preferences rather than expectations when discussing psychological mechanisms (e.g. Loewenstein and Haisley 2008, Heilmann 2014).

The goals of this paper are thus first to broaden the theoretical framework used to describe mechanisms underlying nudges along these dimensions, and then explore some normative implications of this extended framework. Specifically, I will focus on *social nudges* that aim at increasing people's voluntary provision of public goods. Social nudges are an ideal case for my purposes because this type of nudge exploits both individual and collective mechanisms that do not figure in more famous examples of nudges. To set the stage I will begin by clarifying normative issues before discussing how social nudges work. First I will characterize *paternalism* in general, and *nudge paternalism* in particular, explaining what is distinct about the criticisms of the latter. I will identify two objections, *the objection from autonomy* and *the objection from coherence*. I shall argue that the use of social nudges can be justified against both objections, but for different reasons.

## 2 Normative Critiques of Nudge Paternalism

What is *paternalism*? As Dworkin (2013) notes, the answer to this conceptual-normative question depends on the context and goal of analysis. Our context is that in which the moral acceptability of a new form of paternalism, nudge paternalism, is in question; and our goal is to set a framework in which different positions in this debate can be sorted. To do this I propose the following definition: first of all, paternalism is predicated on some form of *interference* on someone's decision making. How "active" interference should be to count as paternalistic is debatable. A wide view is that all decisions are situated in particular contexts, and therefore any change of contexts counts as interference. In contrast a narrow view requires more direct forms of interference. For our purposes the wide view is more appropriate, and indeed it seems to be the one adopted by Thaler and Sunstein (2008). Second, we can distinguish the means and ends of interference. The goal of the paternalizer (call her X) is *beneficence*, namely to make the paternalizee (call him Y) better off in some sense, e.g. financially, psychologically, or physically. The means is not specified by the notion of paternalism *per se* (Dworkin 2013, fn. 25), although it is typically the case that the measures used to benefit Y exploit some asymmetric relation between X and Y, e.g., X having more resources, knowledge, power, deliberative skills, etc., than Y does (thus we do not talk about children acting paternalistically towards their parents, or citizens towards their state). This focus on means and ends

helps us to see that some debates concern whether the goal of paternalism itself is justified, while others concern whether the means are justified, independently from the goal.

The issue of whether the goal (beneficence) is justified pertains not only to nudge paternalism but to paternalism in general. For this reason I shall not discuss it here (see Sugden 2009, Grüne-Yanoff 2012). Another distinct critique against nudge paternalism concerns nudges as a specific means of paternalism. Thaler and Sunstein (2008, 6) characterize a nudge as follows:

> A nudge, as we will use the term, is any aspect of the choice architecture [i.e., the context in which people make decisions] that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid.

This negative characterization of a nudge is intended to highlight its appealing feature that it does not significantly limit the individual freedom of choice. To use the famous examples from Thaler and Sunstein (2008), it is supposed to be easy and cheap to avoid healthy food in a cafeteria, or opt out from a contribution plan for retirement saving. And yet the choice architect (X) can increase the likelihood of Y's choosing healthy food or wise saving plan simply by changing subtle features of the choice architecture. Let us suppose for the moment that in each case X's goal of beneficence is uncontroversial (so imagine that Y will be better off "nudged"). A number of philosophers argue that these types of nudges may still diminish liberty in a wider sense than the freedom of choice, namely through interference with one's self-governance or *autonomy*—'the control an individual has over his or her own evaluations and choices' (Hausman and Welch 2010, 128). Such nudges, philosophers complain,

> 'deliberately circumvent people's rational reasoning and deliberating faculties, and instead seek to influence their choices through knowledge of the biases to which they are susceptible' (Grüne-Yanoff 2012, 636 )

> exploit 'flaws in human decision-making to get individuals to choose one alternative rather than another' (Hausman and Welch 2010, 128)

Since these quotes seem to rely on different intuitions about why nudges are autonomy-damaging, it is useful to explicitly distinguish relevant notions of autonomy here (for a thorough review of alternative notions of autonomy see Buss 2014, Section 2). First, terms like "flaws" and "biases" assume some rational ideal of choice behavior from which individuals deviate. In economics this ideal is modeled as certain consistency conditions in choice, such as transitivity of preferences. This is an instance of the general notion of autonomous agency as the capacity to base actions on internally coherent mental states (*the coherentist view*). On this view, nudges are autonomy-damaging when they cause inconsistent preferences or more generally incoherent mental states. Call this *the objection from coherence*. Second, verbs such as "circumvent", "influence", and "get individuals to choose" suggest a violation of

autonomy in terms of reasoning *processes*. The underlying intuition corresponds to the notion of autonomous agency as the capacity to reason and respond to reasons (Buss 2014).[1] Nudges, according to this intuition, are morally unacceptable when they induce behavioral changes to which one's reasoning process is not responsive. Call this *the objection from autonomous reasoning* or *autonomy* for short.

In what follows I will examine both objections in turn, and show that there are plausible mechanisms through which social nudges work that is not autonomy-damaging in the senses explicated in both objections.

## 3 The Objection from Coherence

In the rational choice tradition, the concept of autonomy has been discussed in terms of the rationality of preference change because, in this tradition, behavioral change is theoretically closely tied to change in preference. The crucial question thus becomes whether a particular preference change is rational in some specified sense. In criticizing Elster's (1983) theory of rational preference change qua *intentional* preference change, Bovens (1992) proposes his alternative, according to which a preference adjustment is rational only if it is informed by the agent's overall preference structure, all-things-considered. For example, the fox in La Fontaine's fable, who changes his preference for the grapes because they are out of reach, is irrational because he changes this particular token preference for those grapes in an *ad hoc* way, without changing his type preference for the sweet, juicy summer fruits (a case of sour grapes, or SG). In contrast, imagine that I lost the current postdoc position and had to take a research administration job. Suppose I, as a result of this career choice, start to better appreciate my life in general, by enjoying the possibility of planning my life over a longer term (having a family, start saving, buying a house, etc.) and reading more fiction at home, etc., and say "Oh how I love my job more than the postdoc." Then my preference change for the current career over the academic one is not SG but rational, because my global preferences have also changed in such a way that this particular preference coheres with them. In other words, now I am the kind of person who enjoys this kind of career with more stability and more time for leisurely reading (a case of character planning, or CP). This theory (Bovens 1992) says that a rational choice is one that coheres with the agent's overall, all-things-considered preferences.

Based on this theory, the objection from coherence can be stated more precisely: nudges are morally problematic to the extent that they induce preference change that disturbs the overall coherence of the all-things-considered preferences of the agent. Bovens (2009) applies this criterion by first distinguishing six different types of nudges based on their target mechanisms, and then criticizes two types which tend to do this (see also Heilmann 2014).

[1] Until Section 4.2 where it becomes relevant I defer the discussion of the distinction between the responsiveness-to-reasons view and the responsiveness-to-reasoning view, which is that the former is strongly externalist (i.e. reasons should correspond to external reality) while the latter is weakly so (i.e. reasoning should respect certain objective norms, but it can accommodate false beliefs) (Buss 2014).

The first four types are acceptable because these nudges try to '[steer] us in the direction of what we consider to be in line with our overall preference structure.' (Bovens 2009, 213) For example, suppose that Jack wants to adopt a dietary regime that is optimal for his health, all things considered. In such a situation, nudges might change Jack's unhealthy dietary choice due to the lack of expertise in nutritional science (Ignorance), forgetfulness or laziness (Inertia), temptations of sweet, salty and fatty foods (Akrasia), or emotional costs associated with thinking seriously about what he puts into his mouth (Queasiness).[2] In these cases, nudges aim to ameliorate a particular local inconsistency so that the preference in question will be adjusted to cohere with Jack's overall preference for healthy dietary choice.

The other two types are more problematic because, in contrast to the first four, these interventions nudge people 'in the direction of agency which [they] do not believe to be in [their] interest' (Bovens 2009, 212), all things considered. *Exceptions* are nudges that target the statistically representative population, at the cost of the rest. For example, the layout change in the cafeteria might nudge people to eat more healthy food, and this might be good for most people who would regret not having eaten healthy food after being sent for a coronary artery bypass surgery. But such nudges might not be good for those exceptional people who, all things considered, genuinely prefer the current unhealthy food choice plus the risk of diseases and earlier death, to the current healthy food choice plus the prospect of living healthier and longer.[3] The same problem arises with many nudges that try to change individual economic behavior (e.g., the default change of the pension scheme to encourage saving), as long as those nudges target a population that is heterogeneous with respect to the overall preferences. Although Thaler and Sunstein's free choice condition intends to prevent such exceptional individuals from getting influenced by nudges, the condition is not generally satisfied because people with exceptional preferences are not always exceptionally resistant to the influence of nudges. This is not a decisive argument against Exception nudges, but nudge paternalists need to provide justification for prioritizing the majority preferences over those of minorities.

*Social Benefits* also try to nudge people to make a choice which they do not believe to be optimal given their individual preferences. The aim is to benefit society at large, instead of the majority of citizens as in the case of Exceptions. Theoretically speaking, Social Benefits, or what Thaler and Sunstein (2008) call *social nudges* (I adopt the latter term hereafter), encourage the voluntary provision of public goods, which are characterized by non-rivalry (one's consumption doesn't reduce that of others) and non-excludability (it is impossible or difficult to exclude people from benefiting from the goods). The standard economic analysis shows that voluntary provision of public goods remains sub-optimal because self-interested individuals can do no better than simply free-riding on others' contributions. This is the situation called a social dilemma. Standard policy recommendations are privatization of a good or

[2]These examples are mine.

[3]They might also have exceptional beliefs that their lives or the world will end before they start to suffer from their dietary choice.

taxation to finance it publicly, or to price a bad. Social nudges in contrast try to increase (decrease) voluntary provision of public goods (bads) by other means.

Given this view of what social nudges do, the objection from coherence states that they are problematic because if they work people will act pro-socially, which is inconsistent with their own self-interested preferences.[4] There are several ways to respond to this objection. First we can point out that most people would endorse the state in which public goods are sufficiently provided, given their overall interests. In a n-player ($n \geq 2$) social dilemma, each player's preference ordering is defined roughly as DC (I defect, others cooperate) $\succ$ CC (I cooperate, others cooperate) $\succ$ DD (I defect, others defect) $\succ$ CD (I cooperate, others defect). Social nudges push an individual against her self-interest in the sense that they nudge her to choose C though she prefers DC to CC. But at the same time, the C-choice enables her to obtain CC, which she prefers to DD. Since DC cannot possibly be the outcome for everyone (that is, if everyone free-rides, there will be no one left to free-ride on), the individual would endorse CC as consistent with her overall interests in the sense that that's a better outcome than DD, if achieved.[5] Of course, this endorsement will be contingent on the fact that the number of the people nudged to the C-choice exceeds a certain threshold[6] such that the payoff of the minimal cooperative outcome is larger than that of the DD for the individual in question.[7]

This response to the objection from coherence becomes even stronger when we drop the purely conventional formulation of the social dilemma and allow that not everyone is narrowly self-interested. Specifically, a bulk of evidence from public goods experiments suggest that a substantial portion of subjects are *conditional* cooperators who cooperate if they expect that a sufficient number of others do the same (for a review see Gächter and Thöni 2007, 188). Since these conditional cooperators do prefer CC to DC, if social nudges create this expectation and achieve CC (or its approximation), then there is no sense in which their C-choices do not cohere with their overall preference structure. Regarding free-riders who complain that they were nudged to C although they really wanted DC rather than CC, there is no practical or ethical reason to respect such preferences, unlike in the case of Exceptions above. As to those free-riders who were not nudged, they have nothing to complain about since they actually became better off.

Another, perhaps more radical way to respond to the objection from coherence is to deny that there is such a thing as the overall preference structure corresponding to the all-things-considered evaluation of the individual.[8] Bacharach et al. (2006, Introduction) explicate this idea as *variable frame theory*. In contrast to the standard

[4]Bovens (2009) doesn't state this explicitly (see footnote 8 on page 214), so this is my reconstruction.

[5]Here I am interpreting the social dilemma as a dilemma *for individuals*, as well as for social planners. Some game theorists adopt an alternative interpretation according to which individuals do not face any dilemma as they simply choose dominant strategies based on their preferences. On this reading a C-choice is against their self-interest by definition, although such a view is not common in social sciences.

[6]This number is Schelling's *k* Schelling (1978, ch.7).

[7]I thank an anonymous reviewer for pointing this out.

[8]I thank an anonymous reviewer for giving me the courage to say this.

literature on framing that presupposes an objectively identifiable set of choices, variable frame theory suggests that (i) a choice set becomes meaningful only after framed in a certain way; (ii) different frames make different objects of choice salient; and (iii) in effect, people choose from different sets under different frames. For instance, ($\Delta$, O, X) can be 'delta, omicron, chi' in a Greek frame or 'triangle, circle, cross' in a geometry frame, but one cannot say that the agent's choice reveals inconsistent preferences if she chooses O in the Greek frame, while choosing X in the geometry frame: the agent is in a way choosing from a *different* choice set in each case. In social dilemma and coordination games, Bacharach et al. (2006) specify two frames that a player may adopt, namely I-frame and we-frame. The standard game theory implicitly assumes that a player in social dilemmas always adopts an I-frame (asking "What should *I* do?"), leading to the dominant reasoning ("whatever others do, I will be better off defecting"). But she could be adopting, argue Bacharach et al. (2006), a we-frame (asking "What should *we* do?"). Players who are in the we-frame will choose C, an action that is part of the strategy profile that uniquely maximizes the group's payoff (CC). If individuals are nudged to adopt the we-frame rather than the I-frame (or *vice versa*), this does not make their preference structure inconsistent because the former follows from and is relative to framing (more on this in Section 4.2).

To sum up, there are plausible mechanisms of social nudges that are not vulnerable to the objection from coherence. Generally speaking, social dilemmas are situations where individual rationality leads not only to social inefficiency but individual inefficiency (i.e., CC is both individually and collectively more desirable than DD), and it is widely recognized that some institutional interventions are legitimate (some even believe that the legitimacy of the state depends solely on solving social dilemmas). From this perspective social nudges are not different from other types of institutions that solve social dilemmas. That said, social nudge paternalists still owe an answer to *the objection from autonomy* since this objection concerns not only *what* social nudges do, but *how* they do it. I now turn to this question.

## 4 The Objection from Autonomy

The objection from autonomy states that nudges are ethically problematic to the extent that they change individual behavior in such a way that is not responsive to the agent's reasoning process. To address the objection we thus need to explicitly discuss how social nudges affect behavior. In responding to the objection from coherence in the last section, I briefly sketched two mechanisms of social nudges, namely (i) the creation of the expectations that others will cooperate among conditional cooperators (the expectation-based mechanism); and (ii) the shift from I-frame to we-frame (the frame-based mechanism). In each case, if the proportion of optimistic conditional cooperators or of team-reasoners becomes large enough, CC can be achieved. In this section, I will discuss an example of successful nudges in order to evaluate the force of the objection from autonomy in a concrete policy context, while developing this sketch of the mechanisms of social nudges. Note that my goal here is not to offer the empirically most plausible explanation of why this particular nudge worked, but rather to develop my responses to the objection from autonomy.

### 4.1 Social Nudges as Social Norm Engineering

In a chapter titled *Following the herd*, Thaler and Sunstein (2008, 60) acclaim "Don't Mess With Texas", a highly successful social advertisement campaign against littering on Texas's highways initiated by the Texas Department of Transportation. Using a variety of media ranging from TV social advertisement featuring Texan celebrities to logo T-shirts and mug cups, the campaign reduced the amount of litter in Texas by 29 % in the first year of the campaign, and in its first six years visible roadside litter was reduced by 72 %. How has this been achieved? Thaler and Sunstein (2008) do not provide an explanation of this specific case, but instead suggest roughly two accounts of how social nudges can encourage prosocial behavior. These are expectation-based and priming-based accounts. In what follows I reconstruct and extend them as the explanations of the success of "Don't Mess With Texas" in order to facilitate the discussion of the permissibility of social nudges. I will discuss the priming-based account in the next subsection, and here reconstruct the expectation-based one, drawing on Bicchieri's (2006) well-known model of social norms.

The expectation-based account suggests that the campaign somehow created the expectation that others will refrain from littering among the Texans who are conditional followers of social norms. Bicchieri (2006) models such an individual as having a *conditional* preference such that she prefers to conform to a certain behavioral pattern on condition that she believes (a) that enough people are conforming (empirical expectation) and (b) that enough people are expecting her to conform (normative expectation). Although this model predicts that people follow a certain social norm when both of these conceptually distinct conditions are satisfied, the two types of expectations are causally connected. First, at the individual scale there is a psychological mechanism which detects normative expectations from empirical ones in certain social contexts. For example, a college student is less likely to engage in binge drinking when informed that alcohol abuse is less pervasive among fellow students than he thinks it is (Thaler and Sunstein 2008, 67). Weakening an empirical expectation (Others are drinking hard) can weaken a normative expectation (Others expect me to drink hard), thereby reducing conformist behavior. Second, this mechanism can operate dynamically at the aggregate scale, since everyone's behavior affects everyone else's by affecting one's empirical expectation. So the first student who stopped binge drinking may affect the next's empirical and normative expectations, thereby influencing her to stop binge drinking as well, and so on. But in order for this social process to unfold, a certain threshold (the so-called critical mass or tipping point, which can take a different value for different people) needs to be reached.

Based on this reconstruction, Thaler and Sunstein's (2008) expectation-based explanation of the success of "Don't Mess With Texas" would be like this: the Texas authority successfully raised the conditional norm followers' expectations that enough others expect them not to litter, with a "tough-talking slogan". Individuals who have stronger conformist preferences (i.e. whose critical numbers of "enough" others are relatively low) change behavior first, which in turn change behavior of those with weaker conformist preferences, until eventually all conditional

norm-followers have changed behavior. This is exactly the opposite case of the dissipation of undesirable norms such as binge drinking among college students.

Does the objection from autonomy carry any weight against social nudges that exploit these mechanisms? In these cases, actions are produced through the right kind of belief-desire psychological process, i.e., people's conditional preference for norm conformity is activated by satisfying two epistemic conditions, namely empirical and normative expectations. In this sense social norm engineering operates through practical reasoning, and therefore is unproblematic. One might argue that it exploits irrationality because an empirical expectation (Enough people are not littering) does not imply, logically speaking, a normative one (Enough people expect me to refrain from littering). Though this is generally true, note that in this specific example there is an additional assumption that littering is illegal and morally condemned (and that this is common knowledge). With this assumption, the inference seems valid, or at least very reasonable.[9] Also norm engineering to some extent respects heterogeneous preferences since it can in principle affect all conditional norm followers with different degrees of conformist preferences, through the social mechanism described above. The recalcitrant nonconformists who cannot be nudged have to be incentivized by conventional means such as fines and imprisonment.

One might still object that these mechanisms operate unconsciously, and therefore outside the realm of practical reasoning (Thaler and Sunstein (2008) seem to worry about this feature of nudges in general). However whether an action was conscious or unconscious is not a relevant criterion for whether that action was delivered through practical reasoning. Although ample research suggests that most of our everyday choices are made unconsciously, we still consider ourselves as the authors of these choices, *post hoc*. What is relevant is rather whether the targeted mechanism can be, conscious or unconscious, characterized as reasoning of our own. I thus conclude that the objection from autonomy is not convincing enough to reconsider the use of social nudges qua norm engineering.

### 4.2 Social Nudges as Framing

Another account of social nudges refers to the frame-based mechanism. Specifically, social nudges may induce people to shift from I-frame to we-frame in social dilemmas, thereby increasing pro-social, group-oriented behavior. Whether and how this frame-shift takes place has been lively discussed by researchers, some suggesting that the increase of voluntary contributions in public goods experiments are mostly due to the change of expectations rather than the shift in framing (e.g. Guala et al. 2013). If this is the case, the response in the previous sub-section will suffice. But since the empirical jury is still out, I will discuss possible mechanisms of frame-shift, and then

[9]Thaler and Sunstein (2008) refer to the "spot-light effect", which makes us believe that others pay more attention to us and our behavior than they really do when we become aware that we look or behave differently from others. Although this effect sometimes creates false beliefs—e.g., that the audience of my presentation scorn my shabby jacket when no one actually cares—this doesn't seem to explain the case of "Don't Mess With Texas" because in a typical Texan road there aren't many people around you.

try to defend social nudges that exploit these mechanisms against the objection from autonomy.

The key question, when considering autonomy as responsiveness to reasoning, is whether a frame-shift is responsive to practical reasoning. If it isn't, nudging people to adopt one frame rather than another doesn't seem to respect their autonomy. If it is, in contrast, a frame-shift can be defended in a similar way as the shift of expectations.

Bacharach's (2006) view on this matter is ambivalent.[10] On the one hand, his view of framing as "an involuntary processing stage which is neither rational nor irrational" (p. 23) suggests that he didn't see it as part of reasoning to which the standards of rationality applies. On the other, he saw team reasoning that solves some coordination and cooperation problems as "good game-theoretic reasoning" (ibid.). In particular, Bacharach's *interdependence hypothesis* proposes that the salience of three features of a game makes it more likely that the players team reason, i.e., identify themselves as members of a team (group identification), frame the choice situation as team members (we-framing), and play their part to achieve the team's goal. The first feature is the existence of *common interest*, defined as the fact that players prefer outcome s* to outcome s. The second feature is *co-power*, the fact that s* can be brought about only by an appropriate combination of their actions, and the third is the existence of a Nash equilibrium that realizes s rather than s*, making the attainment of s* unassured by their individualistic decision-making (see Smerilli 2012). An intuition behind these conditions, roughly put, seems to be that people adopt a we-frame when there is good reason to do so, e.g., when it gives extra benefit to individuals that I-frame cannot deliver. On this view the adoption of a we-frame is responsive to justifiable reasons.[11] Here I offer a two-fold response to the objection from autonomy, without resolving the ambivalent nature of Bacharach's theory of team reasoning. First, if the frame-choice precedes practical reasoning, nudging people to adopt one or the other frame bypasses their practical reasoning. Second, if the adoption of a we-frame tends to occur when it is beneficial to individuals, we can say that the choice of the we-frame is responsive to reasons. Since Bacharach seems to accept both conditional statements, it implies that social nudges that exploit we-framing do respect responsiveness to *reasons*, but not *reasoning*. I did not highlight this distinction in Section 2, but it matters which notion of autonomous agency one adopts, if Bacharach is right in claiming that a frame choice lies outside the realm of practical reasoning.[12]

[10] I thank an anonymous reviewer for making me realize this point clearly.

[11] Bacharach et al. (2006, ch.3) speculate on the evolutionary mechanism of the human propensity to adopt a we-framing in certain coordination and cooperation problems.

[12] Note however that Bacharach's view that framing is neither rational nor irrational is not widely accepted. First of all, it seems to be in tension with his interdependence hypothesis and his model of circumspect team reasoning that explicitly weighs the risk of others not adopting a we-frame. Another prominent team reasoning theorist Sugden (1993) also suggests that for a player to adopt a we-frame, she needs *assurance* or *common reason to believe* that enough others do the same. More generally, many emotion researchers emphasize that the way we construe (or frame) a decision situation is *motivated* by our salient *concerns*, i.e., 'the attachments and interests from which many of our desires and aversions derive.' Roberts (2003, 142).

One might object to the second part of the response by saying that social nudges often try to activate a we-frame even without making any reason salient. That is, a shift from I-frame to we-frame might not even be responsive to reasons in the sense specified by the interdependence hypothesis. This objection questions the reason-relevant connection between social nudges and we-framing. In general, framing of a choice problem can be influenced by contextual factors that bring to mind a particular set of ideas (Bacharach et al. 2006, 12). This effect is called *priming*. Priming uses some words, symbols, etc., which increase the ease with which certain information comes to mind, which in turn stimulates certain action. Thaler and Sunstein (2008) mention certain examples. Objects characteristic of business environment make people less cooperative and more competitive; giving cold coffee makes people perceive others as more selfish and less sociable, compared to when given hot coffee.

Now, I grant that these priming effects are worrisome in terms of the relevance of reasons for action since we become less sure of why we chose the way we did as the intentional tie between the cue and the resulting action becomes weaker. (Does a warm cup of coffee give us a *reason* to think that someone is a warm person?) Although we can often rationalize our judgement and choice by giving some *ad hoc* intentional description under which it is based on some reason, this move tends to inflate the realm of reasons without a principled theoretical motivation (Bovens 1992).

This worry is highlighted in Hausman and Welch's (2010) account of how "Don't Mess With Texas" worked as a social nudge:

> [The campaign] attempted to create a machismo image for those who don't litter [...] and to influence behavior via that association. [...] By playing on emotions that ought to be irrelevant to littering, it attempted some very mild shaping and thereby influenced behavior at a much lower cost than harsher penalties for littering or expanded enforcement of anti-littering laws. (p. 134)

I shall first comment on this specific account, before coming back to a more general issue of the relevance of reasons for framing. Hausman and Welch's assumption that the positive emotions associated with a machismo image 'ought to be irrelevant to littering' leaves unexplained why that particular irrelevant image did get successfully associated with anti-littering. Moreover, they neglect one important aspect of the case, which is also left unexplained by Thaler and Sunstein's account. Thaler and Sunstein (2008, 60) cite (from an unspecified source) that the public officials of Texas decided they needed "a tough-talking slogan that would also address the unique spirit of Texas pride". The Texas pride naturally has a relevant connection with littering in the highways because it pollutes the state that Texans are proud of. The machismo image itself may be irrelevant to littering,[13] but to the extent that the former is associated with the pride of being a Texan, it has a relevant association with pro-social, group-oriented behavior such as non-littering.

[13]Except for the semantic association between the slogan ("mess with Texas") and littering ("mess up Texas").

At this stage one might say that priming always makes some *pro tanto* reason salient at the expense of others, and that therefore even if it is relevant, this does not justify letting that particular reason determine final choice. But this objection seems to presuppose that there must be *the* relevant, all-things-considered, choice-determining reason, which is like Bovens's overall coherent preference structure the existence of which I questioned based on variable frame theory in Section 3. Of course the questionable theoretical or empirical status of such a reason doesn't entitle nudgers to prime any reason at whim. In particular, there are cases in which the use of priming seems questionable. For example, in the classic "framing effect" (Tversky and Kahneman 1986), (i) an extensionally identical choice set is described in two different ways, priming loss or gain; (ii) loss and gain frames elicit different reasons, in particular *loss aversion* in the loss frame; (iii) in effect, the decision maker may reverse preferences under two descriptions, although she should be choosing the same way because she faces extensionally the same choice set (the violation of descriptive invariance). Since the decision maker susceptible to this framing effect can systematically lose her resources, we can say that she has no good overall reason to get primed to shift from one frame to the other, even if loss aversion is a kind of reason *pro tanto*. So the use of this kind of framing effect for nudging purposes, though popular in practice, can be questionable.

However, the same criticism doesn't apply to social nudges that prime we-framing. For example, think of priming by game labels: a public goods experiment can be labeled either as the Cooperation Game or the Wall Street Game to increase or decrease the level of contributions, respectively, relative to the control condition. The two treatments do not change the payoff structure of the game, and yet influence the subjects' choices. Unlike in the classical framing effect, though, one would not say that subjects who shift frames violate descriptive invariance and therefore are irrational. Descriptive invariance has a normative force when its violation tends to make decision-makers systematically lose, as in the framing effect in the lottery choice. But in coordination and cooperation games like this, being primed by game labels in fact tends to improve one's payoff, since other players tend to be primed in the same way. In the Wall Street Game, there are less contributions from others, so the D-choice gives the DD-like outcome, which is preferred to CD. While in the Cooperation Game, there are more contributions, so the C-choice gives the CC-like outcome, which conditional cooperators preferred to DC, and which free-riders prefer to DD. Systematically responding to priming in this way seems to be part of good game-theoretic reasoning in the sense that it tends to bring in better payoffs to players, given how other players respond to the priming.

Let us come back to "Don't Mess With Texas" to take stock of my responses to the objection from autonomy. First, there are at least two distinct mechanistic accounts of this campaign's success. One is that it increased the expectations of the conditional norm followers that enough others will not litter and that enough others expect them not to litter, using "the tough-talking slogan" and other means. I argued that this social nudge is permissible primarily because the mechanism does not damage autonomous agency in the sense of the capacity to reason. Another account is that the campaign

primed "the unique spirit of Texas pride", which nudged some Texans to adopt a we-frame, in which littering was seen as "the problem of we Texans", and those Texans played their part to solve this problem by incurring individual costs of keeping litter in their cars. I argued that priming a we-frame in this context is permissible if autonomous agency is interpreted as the capacity to respond to non-arbitrary reasons, although it may be problematic if (a) autonomy is interpreted as the responsiveness to reasoning, and (b) framing is a neither rational nor irrational process that takes place prior to reasoning.

## 5 Conclusion

In this paper I have argued that social nudgers can deal with the objections from coherence and autonomy, with necessary qualifications. Specifically, I showed that conditional cooperators or even free-riders may endorse their nudged pro-social behavior given their overall preference, and that there is a way in which social nudges encourage pro-social behavior by priming good reasons. My arguments draw on two mechanistic explanations of how social nudges work, one based on expectations and the other on frames. Also, I emphasized that these mechanisms take place at both individual-psychological and collective-social scales. I hope these explanations will complement the literature's rather narrow focus on individual cognitive mechanisms of nudges. As a concrete example of social nudges, I picked up the "Don't Mess With Texas" campaign and discussed in some detail how it worked. Of course the campaign in reality is not a single intervention but a bundle of different strategies with elements of both nudges and more traditional regulations. So my claim is not that these two accounts exhaust all possible mechanisms. Also given the heterogeneity of the Texan population and of contexts in which they made choices, multiple psychological and social mechanisms must have been at work. Rather, my point was to sketch, by way of discussing this salient example, a theoretically and empirically motivated way of defending the use of social nudges in general against the two serious moral objections to nudge paternalism.

**Acknowledgments** The idea of this paper was born in Manchester where I worked as Sustainable Consumption Institute postdoc for the project *Motivations of Indifference* led by late Peter Goldie. I regret that I cannot show this to Peter. I thank Michael Scott for his support during the project. The paper developed thanks to the comments I received in Manchester, Rotterdam, Trento, Helsinki, Oviedo and Madrid. I am grateful to the University of Oviedo for the research visit grant, and Armando Menéndez Viso for his hospitality. I also thank David Teira for inviting me to UNED, Madrid. Two anonymous reviewers and the editors helped me improve the paper substantially, for which I am grateful. Finally I thank Miles MacLeod for a "buddy" language check. All the remaining mistakes are mine.

**Compliance with ethical standards** This research was initiated while I was funded by Sustainable Consumption Institute at the University of Manchester (2009-2010); I am currently funded by TINT/Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences, University of Helsinki. The paper also benefited from my research visit to the University of Oviedo (Nov. 2014) funded by the same university. I believe there is no conflict of interest involved in this research.

## References


Bacharach, M., N. Gold, and R. Sugden. 2006. *Beyond individual choice: teams and frames in game theory, Princeton*. N. J.: Princeton University Press.

Bicchieri, C. 2006. *The Grammar of Society*. Cambridge, England: Cambridge University Press.

Bovens, L. 1992. Sour grapes and character planning. *The Journal of Philosophy*, 89(2): 57–78.

Bovens, L. 2009. The ethics of nudge. In *Preference Change: Approaches from Philosophy Economics and Psychology,* chapter 10, eds. T. Grüne-Yanoff, and S. O. Hansson, 207–219. Springer, New York.

Buss, S. 2014. Personal autonomy. In *The Stanford Encyclopedia of Philosophy* (Spring 2014 ed.), ed. E. N. Zalta. Stanford University.

Camerer, C., S. Issacharoff, G. Loewenstein, T. O'Donoghue, and M. Rabin. 2003. Regulation for conservatives: Behavioral economics and the case for 'asymmetric paternalism'. *University of Pennsylvania Law Review*, 151(3): 1211–1254.

Dworkin, G. 2013. Defining paternalism. In *Paternalism: theory and practice,* Chapter 1, eds. C. Coons, and M. Weber, 25–55. Cambridge University Press, UK.

Elster, J. 1983. *Sour Grapes*: Cambridge University Press.

Gächter, S., and C. Thöni 2007. Rationality and commitment in voluntary cooperation: Insights from experimental economics. In *Rationality and Commitment,* Chapter 8, eds. F. Peter, and H. B. Schmid, 175–208. Oxford University Press, Oxford.

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

Grüne-Yanoff, T. (2014, Februrary). Why behavioural policy needs mechanistic evidence. unpublished manuscript.

Guala, F., L. Mittone, and M. Ploner. 2013. Group membership, team preferences, and expectations. *Journal of Economic Behavior and Organization*, 86: 183–190.

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

Heilmann, C. 2014. Success conditions for nudges: a methodological critique of libertarian paternalism. *European Journal for Philosophy of Science*, 4: 75–94.

Loewenstein, G., and E. Haisley 2008. The economist as therapist: methodological ramifications of 'light' paternalism. In *The Foundations of Positive and Normative Economics: A Handbook, The Handbooks in Economic Methodologies,* Chapter 9, eds. A. Caplin, and A. Schotter, 210–245. Oxford University Press, Oxford.

Roberts, R. C. 2003. *Emotions: an essay in aid of moral psychology.* Cambridge. UK: Cambridge University Press.

Schelling, T. C. 1978. Micromotives and Macrobehavior, W.W. Norton & Co, New York.

Shafir, E. 2013. *The behavioral foundations of public policy, Princeton*: Princeton University Press.

Smerilli, A. 2012. We-thinking and vacillation between frames: filling a gap in bacharach's theory. *Theory and Decision*, 73(4): 539–560.

Sugden, R. 1993. Thinking as a team: Towards an explanation of nonselfish behavior. *Social Philosophy and Policy*, 1: 69–89.

Sugden, R. 2009. On nudging: A review of nudge: Improving decisions about health, wealth and happiness by Richard H. Thaler and Cass R. Sunstein. *International Journal of the Economics of Business*, 16(3): 365–373.

Thaler, R. H., and C. R. Sunstein. 2008. *Nudge: improving decisions about health, wealth and happiness*. New Haven: Yale University Press.

Tversky, A., and D. Kahneman. 1986. Rational choice and the framing of decisions. *The Journal of Business*, 59(4): S251—S278.