From Libertarian Paternalism to Nudging—and Beyond

Adrien Barton [1,*]

Email adrien.barton@gmail.com

Till Grüne-Yanoff [2]

[1] Institute of Scientific and Industrial Research (I.S.I.R.), Osaka University, Osaka, Japan

[2] Avdelningen för Filosofi, KTH, Stockholm, Sweden

1. Introduction

How can the government influence people to make better decisions about health, wealth and happiness without coercing them? This question has motivated legal scholar Cass Sunstein and economist Richard Thaler to propose a set of policies under the umbrella term of "nudges" (Sunstein and Thaler 2003; Sunstein 2013; Sunstein 2014; Thaler and Sunstein 2008). Largely drawing on the psychological research in decision-making, nudges aim at influencing people to make better decisions, while leaving intact their freedom of choice. Considering people as organ donors by default, changing the shape of plates to reduce calorie intake, framing risks about medical treatments, reminding people vividly about the health consequences of smoking, arranging canteens so that consumers would chose healthier dishes, using "tough-talking" slogans like "don't mess with Texas": all those nudges aim at attaining desirable outcomes without coercion. This proposal has raised a remarkable echo amongst policy makers, with the foundation of a "Behavioural Insights Team" in the UK, and similar organisations by various other administrations.

Since its inception, the idea of nudging has been heavily debated, and experts from various fields have investigated the epistemological, ethical and legal underpinnings of nudge policies. A significant part of the discussion has focused on so-called "libertarian paternalism", which aims at justifying the governmental use of nudges

that benefit the nudgee. But nudges extend beyond libertarian paternalism, and the present volume investigates some of the issues they raise. It gathers contributions by philosophers, psychologists, economists, neuroscientists and legal scholars, beginning with an invited article by Gerd Gigerenzer, and concluding with a commentary on all articles by Cass Sunstein.

This editorial will present the main philosophical debates raised by nudges, and locate the import of the different contributions to this volume in these debates. We will address successively three topics: the definition of key concepts; the normative justification of nudging policies; and their evidential support.

2. Choice Architecture, Nudge, and Libertarian Paternalism: Definitions

Many of the debates addressed in this volume depend on the details of how one understands nudges. In this section, we therefore review some definitions of "nudge", "choice architecture" and "libertarian paternalism" and propose some clarifying distinctions. In particular, we characterize the choice architecture as the context in which people make decisions, and nudges as interventions on the choice architecture with the aim of steering people's behaviour into specific directions. Finally, we characterize libertarian paternalism as a particular kind of advocacy of nudges.

Thaler and Sunstein (2008) define "choice architecture" as the context in which people make decisions, and a nudge as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives." They add that to "count as a mere nudge, the intervention must be easy and cheap to avoid". Thus, since its inception, nudge has been given a dual nature: as an aspect of choice architecture, or as an intervention. Consequently, nudge has sometimes been defined as a non-intentional entity, and identified with choice architecture (see Mills 2015, this issue); and sometimes considered as having an intentional component, and defined as a tool (Gigerenzer 2015, this issue; Lepenies and Malecka 2015, this issue) or an intervention on the choice architecture that would alter people's behavior (cf. Guala and Mittone 2015, this issue, or Nagatsu 2015, this issue). The latter definition—which is in line with Sunstein 2015 (this issue) - I think this should be a long dash. makes a clear distinction between choice architecture and a nudge, and emphasizes

the intentionality, a morally relevant factor (Saghai 2013); as it is the most differentiated notion, we will employ it here.

The goal of a nudge is to "alter people's behavior in a predictable way" (Thaler and Sunstein 2008), to "steer people in particular directions" (Sunstein 2015, this issue). However, nudge should be "easy to avoid"[1] and not forbid any option; as Thaler and Sunstein (2008) say: "We strive to design policies that maintain or increase freedom of choice". These formulation are vague, and open to interpretation. Saghai (2013), for example, interprets maintaining freedom of choice as the combination of preservation of the choice set and the condition of substantial noncontrol, and proposes the following account of a substantially noncontrolling influence: "A's influence to get B to φ is substantially noncontrolling when B could easily not φ if she did not want to φ." As there are various degrees of controllability, there is a continuum among interventions from rational persuasion, which uses reasons (and is thus fully noncontrolling), to coercion, which uses threats (and is therefore fully controlling), with nudges (which are substantially noncontrolling) in-between. Similarly, as a nudge should not "significantly" change the economic incentives, there is a continuum among interventions from nudges on one hand, to full-blown taxes or financial rewards on the other hand: think for example about the "dollar a day" program that pays teenage girls who already have a baby one dollar for each day they are not pregnant (Thaler and Sunstein 2008, p. 234), which lies in-between a nudge and a financial reward. With these clarifications, we can summarize: a nudge is defined here as an intervention on the choice architecture that is predictably behaviour-steering, but preserves the choice-set and is (at least) substantially non-controlling, and does not significantly change the economic incentives.

The design of nudges largely relies on results from behavioural sciences about the use of heuristics—i.e., strategies of judgment or decision that are fast and use only a few cues (instead of the totality of the available information). As a matter of fact, nudges have often been considered as having a special connection with non-deliberative faculties (see e.g., Heilmann 2014 for a detailed account) - the so-called 'system 1' of dual-process theory (cf. Kahneman 2011). Indeed, Thaler and Sunstein (2008) go into great details about how nudges can exploit such cognitive shortcuts in order to affect behaviour. Some nudges, in order to steer behaviour, will exploit this knowledge by seeking to *trigger* the use of certain heuristics. For example, Rebonato (2012) argues that nudges often affect features of the choice

architecture that people would typically think they do not care about (e.g., positions in a list, default options, framing descriptions), rather than those over which people have explicit preferences (e.g., money, status, etc.).

However, Sunstein (2015, this issue) and other authors (see e.g., Hausman and Welch 2010) point that nudges do not necessarily alter people's behaviour by triggering heuristics. First, nudges may also counteract or *block* the detrimental use of heuristics in certain environments (Guala and Mittone 2015, this issue; Mills 2015, this issue) - think about mandatory cool-off periods before a purchase. Second, some nudges may have no special connection with heuristics at all. Sunstein (2015, this issue) indeed states that some tools, like a GPS, do nudge simply by providing information; and as a matter of fact, providing information can alter people's behaviour in a predictable way, if all agents react to the information in a sufficiently homogeneous way. For example, according to Sunstein's broad definition of nudges, a sign informing that a river is dangerous to swim in would count as a nudge, simply because there are enough people who place a high priority on not drowning. Such a broad definition of nudge does not need to rely on any specific psychological theory of heuristics, a topic on which there are significant disagreements, as detailed by Gigerenzer (2015, this issue); however, most commentators have adopted a more narrow definition.

In the following, we accept Sunstein's broad definition of nudges, but distinguish between three different kinds of nudges: heuristics-triggering, heuristics-blocking and informing. Note that a particular intervention might belong to several of these kinds: think for example about shocking tobacco health warnings, which at the same time inform and exploit various cognitive shortcuts like availability heuristics (Barton 2013); or think about the nudge informing students that their peers do binge-drink less often than what they would assume, in order to trigger a conformist heuristics leading to a healthier behaviour (Haines and Spear 1996). Most of the analysis on nudges has focused on the heuristics-triggering kind, which is often considered ethically and politically problematic (e.g., Rebonato 2012; Saghai 2013, as well as Mills 2015; Whitman and Rizzo 2015; Gigerenzer 2015; Nagatsu 2015; Felsen and Reiner 2015 and Lepenies and Malecka 2015 in this issue).

The distinction into heuristics-triggering, heuristics-blocking and informing nudges differentiates them according to the nature of the interventional mechanism (for more on mechanisms, see Section 4.3). Because nudges are interventions with

specific aims—namely steering people's behaviour into specific directions—one can also differentiate nudges according to the nature of these aims. As proposed by Hagmann et al. (2015, this issue), we can identify at least two different kinds of nudges, which influence behaviour in different directions. The first ones are pro-self nudges, which aim at steering people's behaviour in a private welfare-promoting direction. The second family of nudges are pro-social nudges, which aim at steering people's behaviour such as to promote public goods.

Based on these distinctions, we suggest to characterise Sunstein and Thaler's (2003) libertarian paternalism as the advocacy of governmental use of pro-self nudges. The argument driving this advocacy relies on the purported difference of pro-self nudges to the tools advocated by others forms of paternalism, like incentives or commands. In particular, libertarian paternalism is said to be 'libertarian' because nudges arguably do not interfere with the freedom of choice[2]; and 'paternalist' because the interventions in question are 'pro-self' in the sense of aiming to steer people's behaviour in a private welfare-promoting directions. Paternalism is met with various kinds of ethical and political concerns; identifying the use of nudges as 'libertarian' is supposed to distinguish their use from other forms of paternalism and hence constitutes a new justification for their application. As proposed by Guala and Mittone (2015, this issue) libertarian paternalism can be described as a 'welfarism', where the welfare in question is private (but the justification of some pro-social nudges could also be described as a form of 'welfarism', where the welfare in question is public—cf. Sunstein 2015, this issue).

This characterization of libertarian paternalism has sometimes been criticized as too broad (Hausman and Welch 2010): as a matter of fact, some instances of informing nudges would count as libertarian paternalistic, although they would not be categorized as paternalistic, according to most general definitions of paternalism.[3] Thus, the term 'libertarian paternalism' might be considered as idiosyncratic (Guala and Mittone 2015, this issue), although libertarian paternalism and general paternalism share important commonalities. Most heuristics-triggering (and maybe also heuristics-blocking) pro-self nudges, though, might be genuinely called paternalistic.

We now address the normative question raised by libertarian paternalism: is the governmental use of pro-self nudges justified? The next part will give an overview of the different arguments that have been given in favour of, or against, such a

thesis.

3. Normative Justifications

3.1. The Libertarian Paternalistic Justification of pro-Self Nudges

There may exist different kinds of normative justification for nudges; for example, in situations where more freedom-interfering interventions like prohibitions are justified, nudges might also be justified. However, we will concentrate in this part on the question whether the governmental use of pro-self nudge can be ethically justified on the basis of the specificity of the tools that are used—namely, pro-self nudges, rather than financial incentives or commands as in classical paternalism.

For this, it is important to dissociate the different dimensions of nudges. There seem to be little ethical objections against informing nudges, informing being generally considered as ethically unproblematic. And as a matter of fact, the bulk of the literature has concentrated on heuristics-triggering nudges; those are the ones whose normative justification we are going to investigate here.

3.1.1. Nudges and Rationality

Libertarian paternalism points to the purported evidence from psychological research for the systematic irrationality of human decision makers, in order to justify the use of nudges. Because a person B's decision that is marred by irrationality might be considered to not be truly B's decision at all, the usual arguments against paternalism do not apply: "to whatever extent B's apparent choice stems from ignorance, coercion, derangement, drugs, or other voluntariness-vitiating factors, there are grounds for suspecting that it does not come from his own will, and might be as alien as the choices of someone else" (Feinberg 1986). Nudging a systematically irrational agent thus might be justifiable because it helps the agent realize her own will.

When this argument is used to justify heuristics-triggering nudges, it relies on an outcome-oriented, rather than process-oriented (Charland 2014), conception of rationality: what matters is that the choice fits some notion of rationality (typically satisfaction of preferences), irrespective of whether the process by which this choice was produced is rational. As some authors have suggested, this might be a

consequence of the reliance of nudges on as-if models, and the neglect of underlying mental processes (Berg and Gigerenzer 2010; Grüne-Yanoff et al. 2014). On the opposite, heuristics-blocking nudges may be seen as relying on a process-oriented conception of rationality that would remove supposedly irrational applications of heuristics in certain circumstances.

However, one needs to be cautious with this diagnosis of irrationality. Gigerenzer (this ~~volume~~issue) argues that many decisions that appear to be irrational on a superficial analysis are in fact ecologically rational—that is, they lead to adequate outcomes in the appropriate environment.

The case of time-discounting appears to be especially problematic. As detailed by Guala and Mittone (2015, this issue), because of agents' temporal myopia, it is not clear which self is rational: the present one, or the future one? (see also Hill, 2007) This is especially true if agents cannot recognize that their future selves are fully the same persons as there present selves, as argued by Lecouteux (2015, this issue): in such a case, it may not be irrational for them to strongly (i.e., hyperbolically) discount the interests of their future selves.

But the fact that the diagnostic of irrationality is sometimes difficult to establish does not imply that it is always impossible: as emphasized by Sunstein (2015, this issue), one should be wary of the risk of over-generalizing. In some situations like akrasia or addiction, it might be possible to diagnose irrationality (Guala and Mittone 2015, this issue; Barton 2013). There is thus hope that nudges might help, in such cases, to restore people's rationality.

3.1.2. The Restoration of Preferences

Libertarian paternalists, in contrast to other forms of paternalism, take individuals' own preferences seriously. Nudges, so they claim, steer people's behaviour in a private welfare-promoting direction—that is, in agreement with their personal preferences. In particular, libertarian paternalism claims to be not a form of 'ends paternalism', that would impose some goals to the nudgees, but rather a form of 'means' paternalism, which accepts people's goals and aims at steering people's behaviour towards those goals (Sunstein 2014). However, libertarian paternalists also acknowledge that people's preferences are not always well-formed. Consequently, they interpret private welfare as the satisfaction of the agent's true

preferences - those the agents would have in ideal conditions of complete information, no lack of self-control and unlimited cognitive abilities (Sunstein and Thaler 2003).

Various authors have pointed out that in practice, nudge applications rarely follow this interpretation of private welfare as the satisfaction of true preferences (Fateh-Moghadam and Gutmann 2013; Grüne-Yanoff 2012; Rebonato 2012; Rizzo and Whitman 2009). A libertarian paternalistic planner faces a knowledge problem when attempting to identify people's true preferences. Whitman and Rizzo (2015, this issue) elaborate on this argument by suggesting that people's underlying preferences may not even exist in some situations—like the ones in which they manifest temporal discounting, hot-cold empathy gaps, or framing effects. Moreover, Sunstein and Thaler (2003) definition of ideal conditions for true preferences (namely complete information, unlimited cognitive abilities and no lack of willpower) might appear quite ill-defined, as suggested by Sugden (2008)—who argues that even if they could be clarified, it would make them presumably impossible to know for the libertarian paternalistic planner. If people's true preferences do not exist or are unknowable, it seems to undermine the goal of steering the agents' behaviour towards satisfying such preferences.
AQ2

Moreover, even if some nudges may steer people to act according to their true preferences, they would not do so for *all* agents: preferences are heterogeneous in virtually any large enough population. Some people may prefer to eat unhealthy but tasty food over healthy food, or smoke willingly, even when taking into account the risks involved. Because of this hetereogeneity, the libertarian paternalist faces what Nagatsu (2015, this issue) calls the 'objection from coherence' (see also Bovens 2009; Heilmann 2014; Schnellenbach 2012): nudges may push some people towards a behaviour that is not in agreement with their own true preferences.

Sunstein (2013) argue that this problem might be mitigated through personalized nudges—that is, nudges that would push agents in different directions, depending on their own preferences. However, he also acknowledges that this raises some important technical difficulties. Moreover, personalized nudging will require to accumulate enough information about the nudgees' preferences, which raises significant issues linked to privacy: in particular, it will be impossible in many cases, through personalized nudging, to satisfy all the preferences (including the

preferences about privacy) of an agent who has high privacy concerns. More fundamentally, as argued by Kapsner and Sandfuchs (2015, this issue), an agent may not even want the government to know what are his privacy preferences—so any attempt to gather information about them will defeat the libertarian paternalistic goal of respecting people's preferences.

There is thus a general difficulty for nudges designed in a libertarian paternalistic spirit: basically any nudge will steer at least some people's behaviour against their own true preferences, even when they are welfare-promoting for a majority of the population.

3.1.3. The Easy Avoidability of Nudges

A classical answer to this problem (Sunstein 2015, this issue) is that in contrast to some forms of classically paternalistic interventions, nudges are designed to be easily avoidable (Sunstein 2014; Sunstein and Thaler 2003; Thaler and Sunstein 2008) or resistible (Saghai 2013). Thus, people who want to be nudged can just "go with the flow", and agents who have different preferences can easily avoid the nudge. However, to be able to do so, two conditions need to be fulfilled: an external condition, and an internal one.

The external condition is that to be able to avoid a nudge, one needs in a first place to know that one is being nudged, which requires a sufficient degree of transparency of the nudge. As argued by Bovens (2009), to be avoidable, nudges should not only be type-transparent (the general existence of such nudges is made transparent to the nudgee), but also token-transparent (each specific intervention is made transparent to the nudgee). Many nudges satisfy this conditions, but many others do not (see Hansen and Jespersen 2013 for a categorization, and also Lepenies and Malecka 2015, this issue).

The internal conditions have been analysed by Saghai (2013): to be able to effortlessly resist a nudge, the agent should have special cognitive capacities - namely attention-bringing and inhibitory capacities. But then appears a new issue: if people are boundedly rational, does it decrease their capacities (or their inner power to exert such capacities) to effortlessly resist a nudge's pressure? (Hansen and Jespersen 2013). This is a question for cognitive psychology, on which neuroscientific model may have insights to bring (see Felsen and Reiner 2015, this

issue, for considerations on top-down control versus automatic reasoning).

Finally, even the agents who do have such capacities will experience some cognitive or deliberative costs to escape the influence of the nudges (Heilmann 2014): they suffer, so-to-speak, of a "psychic tax" (Loewenstein and O'Donoghue 2006; Schnellenbach 2012). It seems, therefore, that in the best of all cases, libertarian paternalistic nudges will benefit some agents—hopefully a majority of them—but always at the expanse of a minority of them (Bovens 2009; Lecouteux 2015, this issue), if only because of deliberative costs.

Thus, to be avoidable (or resistible), nudges must be transparent enough, and agents need to have the right set of cognitive abilities; and even in this case, nudges benefit some population at the expanse of some costs—at least, deliberative ones—to some other population. Although this does not mean that pro-self nudges are unjustifiable (we will investigate some possible justifications in part 3.2), one cannot simply argue that they are totally innocuous because they benefit most agents and are easily avoidable.

3.1.4. The Unavoidability of Arranging the Choice Architecture

To make the case for nudge stronger, libertarian paternalists turn to the supposed impossibility not to nudge: as repeatedly pointed by Sunstein (e.g., Sunstein 2015, this issue), some kind of choice architecture always has to be put into place. Libertarian paternalism only suggests giving it a "good" direction, which would globally benefit the agents (Sunstein and Thaler 2003; Thaler and Sunstein 2008). There are a few caveats though.

First, although choice architecture cannot always be made neutral, there are cases where it can—see Gigerenzer's (2015, this issue) example of framing. In fact, increasing the neutrality of the choice architecture, through proper education, training or better design, might be a policy goal in itself (Grüne-Yanoff and Hertwig 2015; Gigerenzer 2015, this issue). Therefore, one needs to be cautious, when using this argument, about what are the details of the situation at hand; for sure, it cannot be a blanket justification of all kinds of nudges.

Second, although 'choice architecture' refers to a decisional context and is thus indeed unavoidable, 'nudging', on the other hand, refers to an intentional

intervention, with the goal of influencing the behavior of agents; and such an intentional intervention can sometimes be avoided (Grüne-Yanoff 2012; Hausman and Welch 2010; Rebonato 2012). Therefore, according to some non-consequentialist ethical views, the unintentional influence of a choice architecture chosen with no purpose in mind could be less ethically problematic than the intentional influence of a nudge.[4]

Third, it is not obvious that, in cases where no neutral cognitive architecture can be arranged, agents should be nudged in a welfare-promoting direction - that is, that pro-self nudges should be put into place. Maybe, in some situations, it would be more justifiable to use pro-social nudges, which may not nudge in the same direction as pro-self nudges. Thus, the impossibility to choose a neutral choice architecture does not necessarily justify libertarian paternalism—at best, it would justify nudging. The direction of the nudging needs to be independently justified.

3.2. Alternative Justifications of Nudges

In the last section, we argued that although a libertarian paternalistic justification of nudging might not fail in all situations, it is weakened in a number of cases in which it pretends to apply. Yet the governmental use of nudges might be justified by other arguments. First, standard paternalistic arguments (Arneson 2005; Conly 2012, Coons and Weber 2013) might justify pro-self nudges as legitimate tools alongside more coercive measures. Second, arguments based on the harm principle,[5] or more generally on the concern about externalities (Hill, 2007), might justify pro-social nudges: for example, Guala and Mittone (2015, this issue) argue that nudging agents to put enough money in their pension schemes is more justifiable as a pro-social nudge avoiding externalities than as a pro-self nudge. Third, nudges might be legitimated by democratic processes. Hagmann et al. (2015, this issue) provides insights about people's opinions on several examples of nudges. It should be noted that ethical arguments for libertarian paternalism or pro-social nudges might have a role to play here. For example, we discussed how pro-self nudges can benefit the majority of a population at the expanse of some costs for minorities: such costs may well be judged as quite small in comparison to the benefits by the population, and thus be accepted through a democratic process.[6]

AQ3

3.3. Comparative Assessment of Nudges

A justification of a nudge might be absolute in the sense that it renders this nudge permissible, or even mandatory. Alternatively, one might assume a comparative perspective, and ask whether these or other justifications render nudges preferable to alternative intervention tools. Such alternative interventions include risk-savvy education (Gigerenzer 2015, this issue); boosting, which seeks to improve people's decision-making competences by enriching their repertoire of skills and decision tools or by restructuring the environment such that existing skills and tools can be more effectively applied (Grüne-Yanoff and Hertwig 2015); and classical paternalistic measures, like economic incentives and legal commands or prohibitions.

Such a comparative perspective requires that the policy tools are distinct alternatives, which, as we discussed above, is not always the case: Sunstein (2015, this issue), for example, argueds that education ismight be a kind of nudging. Yet even if they were distinct alternatives, it would often be possible to combine them in application. For example, many experts recommend that tobacco health warnings should be combined with taxes and other strong measures in order to limit smoking prevalence.

In some situations, however, choices between these interventions need to be made. This may be simply because of limitations of time and funds. Alternatively, one intervention may causally cancel the other. For example, when education seeks to increase the competences and skills about when and how to use which heuristics, it may conflict with heuristics-triggering nudges, which induce their unreflected use—to nudge patients into accepting a treatment by framing its benefits as relative probabilities might be less effective once we educated these patients in the skill of relating relative risk reductions to absolute frequencies.

A comparative assessment would contrast nudges with alternative intervention tools along a number of dimensions. We discuss two normative dimensions here, autonomy and transparency. Section 4 continues this perspective by discussing a number of evidential support dimensions relevant for such a comparison.

3.3.1. Autonomy

Worries about autonomy loom large in the philosophical literature on nudges. Hausman and Welch (2010) have argued that nudges infringe on the autonomy.

More specifically, Wilkinson (2013) stated that a nudge is inconsistent with the nudgee's autonomy if it is manipulative,[7] and the target has not consented to it. Kapsner and Sandfuchs (2015, this issue) argue that insofar as nudges infringe on privacy, they also infringe on autonomy. Bovens (2009) raises the worry that nudges may damage the capacities of reflection, and thus infringe the autonomy of decision-makers in the long-run. It must be noticed that even nudges who can foster the autonomy of most agents will decrease autonomy for some other individuals (White 2013). Personalised nudges could theoretically solve this problem, but are still for now largely unrealized (Felsen and Reiner 2015, this issue). These concerns suggest that nudges fare less well with respect to autonomy than educational or boosting measures.

Several authors have however claimed that some specific nudges can respect autonomy: Cohen (2013) analyses situations of informed consent (but see Blumenthal-Barby 2013), Barton (2013) examines the case of tobacco health warnings, and Trout (2005) investigates debiasing techniques. In this volume, two articles analyse in details some specific nudges and argue that they can respect autonomy. First, Nagatsu (2015, this issue) investigates pro-social nudges, specifically considering the example of the "Don't mess with Texas" anti-littering campaign. He identifies two mechanisms by which such a campaign may work: by changing expectations that others will not litter (and expect them not to litter), and by stimulating a frame switch, as defined by Bacharach (2006), from an "I-frame" (in which the agent asks himself: "what should I do?") to a "we-frame" (in which he reflects on the question: "what should we do?"). He argues that those two mechanisms are compatible with two important aspects of autonomy: they do not cause incoherent mental states, and they are compatible with a responsiveness to reasons. Second, Mills (2015, this issue) argues that a nudge is compatible with autonomy as long as it satisfies four conditions: being intended to facilitate the nudgee's pursuit of her own goal, having an acceptably low opt-out cost, and satisfying conditions of publicity (see below for an analysis of the condition of publicity) and transparency. Mills explains how personalisable default rules, choice prompts and framed information provision satisfy such conditions. This holds across various definitions of autonomy, whether construed as authentic self-rule, or according to a hierarchical or relational account.

Other arguments explain more generally how nudges can actually improve autonomy. On the base of neuroscientific models, Felsen and Reiner (2015, this

issue) argue that if a nudge counteracts a bias, it may be seen as increasing the autonomy of the decision, because it promotes a choice that is in line with higher-order desires. Sunstein (2013) argues that having to make all relevant choices by ourselves could make us worse off, and even less autonomous: given "limitations of time, interest and concern", a choice architecture that guides us towards good decisions would enable us to freely concentrate on the matters that we deem as the most important for us. On a similar note, Mills (2015, this issue) holds that even if some nudges (e.g., default rules, choice prompts and framed information provision) exhibit epistemic paternalist tendencies, they do not contravene on personal autonomy, as insisting on keeping the agent in a state of full epistemic independence would prevent him from being able to pursue important goals. These arguments suggest that at least in these cases, nudges might be on a par, or even supersede, education and boosting measures with respect to autonomy.

Finally, "autonomy sceptics" raise doubts about the value of the concept of autonomy: Saghai (2013) has argued that definition of the preservation of freedom of choice in terms of autonomy (or liberty) would generate confusion, given the lack of consensus on the meaning of autonomy—and he instead proposes to focus on the conditions of choice-set preservation, and full or substantial noncontrol. On the other hand, some sceptics accept the usefulness of this concept, but suggest that the high respect in which we hold autonomy might be misguided. Sunstein and Thaler (2003) suggest that autonomy could be overriden on consequentialist grounds in some settings, and Verweij and Hoven (2012) have argued that many public health policies that restrict choice involve only minor restrictions to personal autonomy. More extremely, Sunstein (2013) suggests that autonomy concerns may actually simply function as a welfare-improving heuristic; on this view, infringement with autonomy may not be a serious problem, as long as we are guided towards welfare-promoting ends. Felsen and Reiner (2015, this issue) argue that neuroscientific evidence suggests that most everyday decisions are not free from undue external influence, which are often integrated in our decision processes in a covert way. Thus, most of our everyday decisions would not count as autonomous according to some conceptions of autonomy, and they ponder whether the high regard in which we hold autonomy may be misguided.

3.3.2. Publicity and Type-Transparency

Another concern about nudge policies is that because of their covert nature, some

nudges are more difficult to scrutinize or monitor (Glaeser 2006). Sunstein (2013) recognizes that some nudges may lack some salience, in the sense that they do not attract the same attention as mandates and bans. We have mentioned in part 2 that if nudges go unnoticed, this constrains the opt-out option; but this also have the effect of undermining social or political resistance. To counter this worry, Thaler and Sunstein (2008) invoke a Rawlsian notion of publicity, according to which policy-makers can only use policies that it would be able and willing to defend publicly to its own citizens. This condition has been criticized in two respects. First, Hausman and Welch (2010) and Lepenies and Malecka (2015, this issue) have raised doubts about its Rawlsian credentials. Second, Lepenies and Malecka (2015, this issue) argue that this condition is insufficient. They explain that nudges are problematic for two reasons: ~~first,~~ most of them are not part of a legal system; ~~second,~~and no nudge—by nature—does impose a requirement to behave in a particular way. Nudges thus rely on an instrumental conception of the law that is problematic on an institutional level. These worries suggest that with respect to transparency, nudges fare less well than legally enshrined incentivising or coercive measures.

As a matter of fact, the publicity condition is hypothetical: it only states that policy-makers can use policies that they *would* be able to publicly defend, but it does not request policy-maker to *actually* defend them publicly. Lepenies and Malecka (2015, this issue) argue that connecting nudges to the legal system would make nudges more visible and accessible,[8] and they suggest various tools in this respect: nudge ombudsman, legal registry of nudge, or information about the legal source of shocking health warnings. Using Bovens (2009) vocabulary, one could analyse such measures as improving the type-transparency of nudges. To summarize, whereas token-transparency is important at the individual level, in order to ensure that the opt-out clause can indeed be chosen (see part ~~2~~3.1.3), type-transparency is important at the institutional level, in order to complement Thaler & Sunstein's condition of publicity.

3.4. Normative Justifications: Conclusion

To summarize, libertarian paternalism is only one of the possible justifications of nudging policies, and not always the less questionable; in particular, as argued by Guala and Mittone (2015, this issue), some nudges that are both pro-self and pro-social might be easier to ethically justify because of their latter effect than because of the former. On the other hand, first survey results suggest that pro-self nudges are

more acceptable to the public than pro-social ones (cf. Hagman et al. 2015, this issue), in which case a democratic justification might have to focus on the nudges' pro-self effects. In any case, to be fully justified, nudges should be compared to possible alternative measures (like education, boosting or classical paternalistic measures) along several criteria that include their transparency - to ensure that the opt-out condition can indeed be chosen, and to enable the society to contest their use - as well as their general effects on autonomy.

We are now going to show that such normative criteria crucially depend on the mechanisms through which nudges operate—a question connected to the larger issue of the evidential support of nudging policies~~, a crucial issue to justify them~~.

4. Evidential Support

Beyond being normative permissible, behavioural policies also must be empirically supported to be justified. In fact, defenders of behavioural policies have often emphasized their evidential support as one of their hallmarks (Sanders and Halpern 2014). Yet what these evidential standards are is often unclear. In the following, we distinguish three kinds of claims that need to be supported by evidence in order to justify behavioural policies: first, that people make systematic biases which can be corrected, or exploited, by an intervention in *some ideal* laboratory environment. Second, that the intervention produces the desired effect in the *actual* target environment—this concerns the external validity of the effectiveness claim. Third, we need information about *how* the intervention produces the desired effect in the target environment—this concerns evidence about the underlying mechanisms. In the following, we review some of the debates around behavioural policies in each of these dimensions.

4.1. Evidential Support in a Laboratory Environment

In most cases, behavioural policies are based on evidence from behavioural experiments. Historically, nudge policies developed out of the *Heuristics and Biases* program. This program started with the goal of showing that actual human judgment and decision under uncertainty diverged systematically from the rational judgment and decision predicted by neoclassical economic models, like probability theory or expected utility theory. The principle method of this program was to perform behavioural experiments in laboratory conditions. Only later did scholars associated

with the *Heuristics and Biases* program employ this systematic divergence claim to justify interventions designed to return people to the rational course of action (for more historical background, see Heukelom 2014), adding a prescriptive layer to the descriptive theory.

Nudges thus are often based on evidence from lab experiments. The first question that arises is whether the Heuristics and Biases conclusions about these experiments are correct. Does actual behaviour in lab experiments systematically diverge from what would be predicted by neoclassical models? And if yes, does it provide evidence for the alternative models proposed by the Heuristics and Biases program? Such questions are heavily debated in economics and psychology departments worldwide.

A first criticism insists that the deviation from neoclassical models identified in some of these experiments are mere artefacts of the experimental design: the experimental results are highly sensitive to changes in the phrasing of the experimental tasks, for example in terms of probabilities of single events rather than frequencies of repeated events (Gigerenzer 1991, 1996; Kahneman and Tversky 1996; Gigerenzer 2015, this issue, X11–X21 The correct page has to be put here.). To illustrate, take the so-called *conjunction fallacy*. Tversky and Kahneman (1983) had shown that subjects, after hearing the description of a person —call her Linda—judge it more probable that the conjunction of two sentences *A&B* are true of Linda than that a single sentence *A* is true of her. Such a judgment of course violates basic probability theory. Gigerenzer (1991), however, pointed out that the purported fallacy disappears if the problem instead is phrased in terms of frequencies. That is, if subjects are told that there are 100 people who fit the Linda description, they then do not judge that there are more people who satisfy *A&B* than only *A*. This raises questions about the Heuristics and Biases interpretation of these experiments as revealing a deeply rooted fallacy (for similar non-robustness results, see Bohm and Lind 1993 and Cubitt et al. 1998).

A second criticism admits that behavioural experiments might provide evidence for the deviation of actual behaviour from neoclassical models, but denies that they provide evidence for any of the alternative models. These critics claim that the alternative models are not sufficiently severely tested: instead of testing their predictive power on out-of-sample experimental data, these models are typically only fitted to the sampled data. Because these models have many degrees of

freedom, it is not difficult, so the critics claim, to reach a high degree of fit with the data, even when the model is misspecified (Berg and Gigerenzer 2010; Binmore and Shaked 2010; see also Binmore and Shaked 2010 and Brandstätter et al. 2008 for a critics of cumulative prospect theory along those lines). This raises questions about the Heuristics and Biases interpretation of these experiments as providing evidence for their alternative models.

4.2. Evidential Support in a Real-World Environment

If the above problems can be overcome, experiments provide support for claims that the intervention produces certain behaviour *in the laboratory environment*. Yet this is not sufficient for policy justification: after all, what matters for the policy is (i) that the bias it is supposed to counteract operates in the target population, and (ii) that the policy intervention produces the desired effect in the target population. For this reason, the *efficacy* of a policy intervention - long dash? - that is, its effect on behaviour under experimental conditions - long dash? - is distinguished from its *effectiveness*, defined as its effect in the intended target population. The possible difference between efficacy and effectiveness constitutes the so-called external validity problem:

> "Efficacy is no evidence whatsoever for effectiveness unless and until a huge body of additional evidence can be produced to show that efficacy can travel, both to the new population and to the new methods of implementation" (Cartwright 2009, 133).

External validity is a general problem for all policies that are proposed based on laboratory experiment evidence (Hogarth 2005). Our impression is that currently, this general problem is not sufficiently addressed when evaluating nudge policy proposals.

Ways to investigate external validity consists either in providing arguments why efficacy in an ideal environment would imply effectiveness in the real environment, or trying to establish validity claims by performing (field) experiments directly in the target population. The first strategy takes recourse for example to mechanistic information (Steel 2008) or information about the relevant similarity of experiment and target populations (Guala 2005). For example, Camerer (2004) argues that

there are important similarity between “naturally occurring field data” and the experimental results, thus providing evidence that supports prospect theory “in the wild”.

Experimental economists have also levelled an important criticism against the application of the Heuristics and Biases Program in the economic domain, and hence against nudge policies. Binmore (1999) argues that economic agents face important decisions repeatedly, and therefore learn how to behave rationally - hence the results from single-shot economic experiments under lab conditions do not apply. Although systematic fallacies of inexperienced subjects might be very relevant for e.g., models of consumer choice, the domain of genuine economic decisions is relevantly different from that of the typical Heuristics and Biases experiments, and therefore conclusions from the experiments cannot be applied to this domain.

Another strategy to overcome external validity problems is to avoid inferences from efficacy to effectiveness altogether and instead rely on field experiments. Such field experiments, so its supporters claim, yields evidence that supports effectiveness straight away and thus does away with the external validity issue (Levitt and List 2009). For example, a recent policy proposal to improve tax collection rates has been tested in two large natural field experiments on the UK population (Hallsworth et al. 2014).

Cass Sunstein, in his closing comments in this special issue, seems to follow this line of argument. He suggests that “even if we … conclude that Gigerenzer is correct on some or all of the psychological issues, we will hardly cease to be favorably disposed toward sensible default rules and good choice architecture.” (Sunstein 2015 this issue, X12 The correct page has to be put here.). That is, even if one doubted the Heuristics and Biases interpretation of the relevant experiments, one should still implement “sensible” and “good” behavioural interventions. But how do we find out that they are “sensible” or “good”? Presumably by testing their effectiveness right in those populations in which they are supposed to be implemented.

Despite the obvious advantages of field experiments over lab experiments when it comes to external validity, problems remain. For one thing, the external validity problem is avoided by field experiments only if the population in the experiment

and the target population are actually the same. The external validity problem reappears if evidence from e.g., a field experiment in the UK is used to justify e.g., a policy intervention in Germany (Cartwright and Hardie 2012). It also reappears when there are considerable time differences between field experiments and the implementation of the policies in that same population. Because during such time periods, a population typically undergoes a lot of changes (e.g., demographic, political, technological), the external validity issue reappears here too.

4.3. Mechanisms

As we mentioned in the previous section, one strategy to overcome the external validity problem is to summon mechanistic evidence. By mechanistic evidence, we mean evidence supporting claims about the underlying mechanism that produces an effect. For behavioural policies, such mechanisms will often be psychological mechanisms—describing cognitive processes that lead from the policy intervention through the agent's perception, feelings, cognition, to her forming an intention and instantiating behaviour. As some of the authors of this special issue argue, neurophysiological mechanisms (Felsen and Reiner 2015 this issue) and social mechanisms (Nagatsu 2015, this issue) also are relevant for assessing behavioural policies. More generally, mechanisms of interaction between the nudges and the environment may also be relevant—for example, does the nudge record as input private information? (Kapsner and Sandfuchs 2015, this issue). Note that mechanistic evidence is evidence for a different thing—i.e., for mechanistic models—than evidence for the effectiveness of an intervention in a particular environment (Illari 2011). The issues discussed in this section therefore cannot simply be remedied by providing better evidence for effectiveness.

Such mechanisms help support inferences from efficacy to effectiveness. Steel (2008), for example, maintains that mechanistic evidence indicates how an intervention works in the ideal experimental situation, *and* to what extent the same mechanisms are also operative in the target population. The latter information then allows us to judge whether an efficacious intervention is likely to be effective in the target population.

Although helpful for inferences from efficacy to effectiveness, mechanistic evidence may seem unnecessary when we have at disposal field experiments, which might obviate the need for such inferences. However, mechanistic information is actually

indispensible for devising such field experiments (Grüne-Yanoff 2015b), for several reasons.

A first reason concerns the robustness of policies. Consider the example of disclosure of conflict of interest: because doctors who experience a conflict of interest tend to treat patients differently than without such a conflict, policies have been proposed that would make the disclosure of such conflicts of interest mandatory. However, although disclosure of conflict of interest successfully inform people's deliberations, they also trigger two other processes that lead to *increased*, instead of the expected *reduced* compliance (Sah et al. 2013): the *insinuation anxiety* lets advisees fear that rejecting advice may signal to the advisor that they believe the advisor is corrupt; and the *panhandler effect* lets advisees feel the pressure to help advisers obtain their personal interests once the adviser discloses this interest (Cain et al. 2011). To understand how the policy is sensitive to such side effects, mechanistic evidence about their operation must be available. Yet to even design field experiments to test the effectiveness of the intervention, a minimal understanding of these sensitivities is required.

A similar problem arises with persistence. A policy is not persistent for example when its effect wears off with time. Yet to understand when to expect such wear-off effects—and how long to run field experiments for in order to test for such effects—evidence about the mechanisms that produce the behavioural effects is required (Grüne-Yanoff 2015b).

Perhaps the biggest need for mechanistic evidence, however, stems not from the external validity issue, but from normative concerns. As we discussed in Section 3, it is debated whether nudging policies violate relevant normative criteria such as autonomy, liberty, transparency or informed welfare maximization. A number of contributions in this special issue seek differentiated judgments, arguing that the normative permissibility is conditional on the process through which the nudge intervention affects behaviour. These contributors include Mills (2015, this issue), who argues that whether nudges violate autonomy depends on the particular influence of the choice architecture on behaviour; Lepenies and Malecka (2015, this issue), who argue that nudges undermine self-legislation when they are cognitively not accessible; and Kapsner and Sandfuchs (2015, this issue), who argue that nudges interfere with privacy to the extent that they require access to private information for their implementation. Each condition—how the choice architecture

influences, whether the nudge is cognitively accessible, whether the nudge requires private information—refers to the mechanism through which the nudge operates. Thus, in these cases, the normative admissibility of the nudge depends on mechanistic information. This argument is even more explicit in Felsen and Reiner (2015, this issue), who argue that neuroscience, by identifying how decisions are made by the nervous system, also illuminates how nudges may affect the rationality of a decision, its correspondence to fundamental goals, and the presence of undue external influences—all factors relevant for determining the degree of autonomy of a decision. It is also pursued by Grüne-Yanoff (2015b), who argues that the welfare-improving capacities of a nudge depend on how the nudge contributes to the formulation and realization of the agent's reflected preferences.

Mechanistic evidence is thus important for a number of dimensions along which behavioural policies are assessed. However, most proponents of nudge currently discuss only candidate mechanistic models, like for example the possibility that the default effect might be produced either through cognitive costs, loss-aversion or a recommendation effect (Smith et al. 2013). Thus, the problem is not that there are no mechanistic models associated with these policy proposals, but rather that there are too many—and that there is hardly ever any evidence provided to choose between them.

5. Conclusion

The academic literature on nudging has followed a process of progressive differentiation: earlier blanket criticisms or defences of nudging policies have progressively been refined and particularized to specific policies. Nudges can vary in several dimensions—both in the cognitive process they entail (heuristics-triggering, heuristic-blocking, or informing) as well as in their goals (pro-self or pro-social)—and such distinctions are critically important to their normative justifications. Even nudges belonging to the same category may raise different normative issues, depending on various factors such as their interaction with autonomy, their institutional transparency, and the mechanisms on which they rely. This implies that while some nudges can be justified via libertarian paternalistic approaches, in other cases justificatory strategies such as standard paternalistic arguments, the harm principle, or democratic decision processes are more promising. Finally, mechanistic evidence will be crucial for supporting the applicability of nudges in real-world settings, and refining their normative

justification.

Acknowledgments

We would like to thank Christophe Heintz for his very careful guidance during the editorial process and his helpful comments on this article, Paul Egré, as well as Stefan van Dijl, Allan Nebres and Edelyn Pervera from Springer for their assistance. Many thanks are due to the reviewers of the articles included in this volume, for the quality of their comments and their swiftness. For financial support, Till Grüne-Yanoff thanks the Swedish Research Council (diarienr. 2011–1302), and Adrien Barton the Japanese Society for the Promotion of Science.

References

Arneson, R.J. 2005. Joel Feinberg and the justification of hard paternalism. *Legal Theory* 11(03): 259–284.

Bacharach, M. 2006. *Beyond individual choice: teams and frames in game theory*. Princeton: Princeton University Press.

Barton, A. 2013. How tobacco health warnings can foster autonomy. *Public Health Ethics* 6(2): 207–219.

Berg, N., and G. Gigerenzer. 2010. As-if behavioral economics: Neoclassical economics in disguise? *History of Economic Ideas* 18(1): 133–166.

Binmore, K. 1999. Why experiment in economics? *The Economic Journal* 109(453): 16–24.

Binmore, K., and A. Shaked. 2010. Experimental economics: where next? *Journal of Economic Behavior & Organization* 73(1): 87–100.

Blumenthal-Barby, J.S. 2013. On nudging and informed consent—four key undefended premises. *The American Journal of Bioethics* 13(6): 31–33.

Bohm, P., and H. Lind. 1993. Preference reversal, real-world lotteries, and lottery-interested subjects. *Journal of Economic Behavior & Organization* 22(3):

327–348.

Bovens, L. 2009. The ethics of nudge. In *Preference change*, ed. T. Grüne-Yanoff and S.O. Hansson, 207–219. Netherlands: Springer.

Brandstätter, E., G. Gigerenzer, and R. Hertwig. 2008. Risky choice with heuristics: reply to Birnbaum (2008), Johnson, Schulte-Mecklenbeck, and Willemsen (2008), and Rieger and Wang (2008). *Psychological Review* 115(1): 281–289.

Cain, D.M., G. Loewenstein, and D.A. Moore. 2011. When sunlight fails to disinfect: understanding the perverse effects of disclosing conflicts of interest. *Journal of Consumer Research* 37(5): 836–857.

Camerer, C.F. 2004. Prospect theory in the wild: Evidence from the field. In *Advances in behavioral economics*, ed. C.F. Camerer, G. Loewenstein, and M. Rabin, 148–161. Princeton: Princeton University Press.

~~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.~~

AQ4

Cartwright, N. 2009. Evidence-based policy: what's to be done about relevance? *Philosophical Studies* 143(1): 127–136.

Cartwright, N., and J. Hardie. 2012. *Evidence-based policy: A practical guide to doing it better*. Oxford: Oxford University Press.

Charland, L. C. 2014. Decision-making capacity. In *The Stanford Encyclopedia of Philosophy (Fall 2014 Edition)*. Edward N. Zalta. http://plato.stanford.edu/archives/fall2014/entries/decision-capacity/ . Retrieved 16 avr 2015.

Cohen, S. 2013. Nudging and informed consent. *The American Journal of Bioethics* 13(6): 3–11.

Conly, S. 2012. *Against autonomy: Justifying coercive paternalism*. Cambridge: Cambridge University Press.Coons, C., and M. Weber (eds.). 2013. *Paternalism: Theory and Practice*, Cambridge: Cambridge University Press.

Cubitt, R., C. Starmer, and R. Sugden. 1998. Dynamic choice and the common ratio effect: an experimental investigation. *The Economic Journal* 108(450): 1362–1380.

Dworkin, G. 2014. Paternalism. In *The Stanford Encyclopedia of Philosophy (Summer 2014 Edition)*. Edward N. Zalta. http://plato.stanford.edu/archives/sum2014/entries/paternalism/ . Retrieved 16 avr 2015.Fateh-Moghadam, B., and T. Gutmann (2013). Governing [through] autonomy: The moral and legal limits of "soft paternalism." *Ethical Theory and Moral Practice*, 17(3), 383–397.

Feinberg, J. 1986. *Harm to self*. Oxford: Oxford University Press.

Felsen, G., and P.B. Reiner. 2015. What can neuroscience contribute to the debate over nudging? *Review of Philosophy and Psychology*. doi: 10.1007/s13164-015-0240-9 .

Gigerenzer, G. 1991. How to make cognitive illusions disappear: Beyond « heuristics and biases ». *European Review of Social Psychology* 2(1): 83–115.

Gigerenzer, G. 1996. On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review* 103(3): 592–596.

Gigerenzer, G. 2015. On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology*. doi: 10.1007/s13164-015-0248-1 .

Glaeser, E.L. 2006. Paternalism and psychology. *The University of Chicago Law Review* 73(1): 133–156.

Goodwin, T. 2012. Why we should reject 'nudge'. *Politics* 32(2): 85–92.

Grüne-Yanoff, T. 2012. Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare* 38(4): 635–645.

Grüne-Yanoff, T. 2015b. Why behavioural policy needs mechanistic evidence. *Economics & Philosophy*, under revision.

Grüne-Yanoff, T. and Hertwig, R. 2015. Nudge versus boost: how coherent are policy and theory? *Minds and Machines*, accepted for publication.

Grüne-Yanoff, T., C. Marchionni, and I. Moscati. 2014. Introduction: methodologies of bounded rationality. *Journal of Economic Methodology* 21(4): 325–342.

Guala, F. 2005. *The methodology of experimental economics*. Cambridge: Cambridge University Press.

Guala, F., and L. Mittone. 2015. A political justification of nudging. *Review of Philosophy and Psychology*. doi: 10.1007/s13164-015-0241-8 .

Hagman, W., D. Andersson, D. Västfjäll, and G. Tinghög. 2015. Public views on policies involving nudges. *Review of Philosophy and Psychology*. doi: 10.1007/s13164-015-0263-2 .

Haines, M., and S.F. Spear. 1996. Changing the perception of the norm: a strategy to decrease binge drinking among college students. *Journal of American College Health* 45(3): 134–140.

Hallsworth, M., List, J., Metcalfe, R., and Vlaev, I. 2014. *The behavioralist as tax collector: using natural field experiments to enhance tax compliance* (Working Paper No. 20007). National Bureau of Economic Research.

Hansen, P.G., and A.M. Jespersen. 2013. Nudge and the manipulation of choice: a framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation* 2013(1): 3–28.

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

Heilmann, C. 2014. Success conditions for nudges: a methodological critique of libertarian paternalism. *European Journal for Philosophy of Science* 4(1): 75–

94.Hill, C.A. 2007. Anti-Anti-Anti-Paternalism. NYU Journal of Law & Liberty 2: 444–454.

Heukelom, F. 2014. *Behavioral economics: a history*. Cambridge: Cambridge University Press.

Hogarth, R.M. 2005. The challenge of representative design in psychology and economics. *Journal of Economic Methodology* 12(2): 253–263.

Illari, P.M. 2011. Mechanistic evidence: disambiguating the Russo-Williamson thesis. *International Studies in the Philosophy of Science* 25(2): 139–157.

Kahneman, D. 2011. *Thinking, fast and slow*. London: Macmillan.

Kahneman, D., and Tversky, A. 1996. On the reality of cognitive illusions, *103*(3), 582–591.

Kapsner, A., and B. Sandfuchs. 2015. Nudging as a threat to privacy. *Review of Philosophy and Psychology*. doi: 10.1007/s13164-015-0261-4 .

Lecouteux, G. 2015. In search of lost nudges. *Review of Philosophy and Psychology*. doi: 10.1007/s13164-015-0265-0 .

Lepenies, R., and M. Malecka. 2015. The institutional consequences of nudging —nudges, politics, and the law. *Review of Philosophy and Psychology*. doi: 10.1007/s13164-015-0243-6 .

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

Loewenstein, G., and T. O'Donoghue. 2006. « We can do this the easy way or the hard way »: negative emotions, self-regulation, and the law. *The University of Chicago Law Review* 73(1): 183–206.
AQ5

Mill, J.S. 1956. *On liberty*. Indianapolis: Bobbs-Merrill.

Mills, C. 2013. Why nudges matter: a reply to Goodwin. *Politics* 33(1): 28–36.

Mills, C. 2015. The heteronomy of choice architecture. *Review of Philosophy and Psychology*. doi: 10.1007/s13164-015-0242-7 .

Nagatsu, M. 2015. Social nudges: their mechanisms and justification. *Review of Philosophy and Psychology*. doi: 10.1007/s13164-015-0245-4 .

~~Oliver, A. 2013. *Behavioural public policy*. Cambridge: Cambridge University Press.~~

AQ6

Rebonato, R. 2012. *Taking liberties: A critical examination of libertarian paternalism*. New York: Palgrave Macmillan.

Rizzo, M.J., and D.G. Whitman. 2009. The knowledge problem of new paternalism. *Brigham Young University Law Review* 9(4): 905–968.

Saghai, Y. 2013. Salvaging the concept of nudge. *Journal of Medical Ethics* 39(8): 487–493.

Sah, S., G. Loewenstein, and D.M. Cain. 2013. The burden of disclosure: increased compliance with distrusted advice. *Journal of Personality and Social Psychology* 104(2): 289.

~~Samuels, R., Stich, S., and Bishop, M. 2002. Ending the rationality wars: How to make disputes about human rationality disappear. In *Common Sense, Reasoning and Rationality* (p. 236–268). Oxford: Oxford University Press.~~

AQ7

Sanders, M., and Halpern, D. 2014. Nudge unit: our quiet revolution is putting evidence at heart of government. *The Guardian*. http://www.theguardian.com/public-leaders-network/small-business-blog/2014/feb/03/nudge-unit-quiet-revolution-evidence . Retrieved 16 avr 2015.

Schnellenbach, J. 2012. Nudges and norms: on the political economy of soft paternalism. *European Journal of Political Economy* 28(2): 266–277.

Smith, N.C., D.G. Goldstein, and E.J. Johnson. 2013. Choice without awareness: ethical and policy implications of defaults. *Journal of Public Policy & Marketing* 32(2): 159–172.

Steel, D. 2008. *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.

Sugden, R. 2008. Why incoherent preferences do not justify paternalism. *Constitutional Political Economy* 19(3): 226–248.

Sunstein, C.R. 2013. Storrs lectures: Behavioral economics and paternalism. *The Yale Law Journal* 122(7): 1826–1899.

Sunstein, C.R. 2014. *Why nudge?: The politics of libertarian paternalism*. New Haven: Yale University Press.

Sunstein, C.R. 2015. Nudges, agency, and abstraction: a reply to critics. *Review of Philosophy and Psychology*. doi: 10.1007/s13164-015-0266-z .

Sunstein, C.R., and R.H. Thaler. 2003. Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review* 70(4): 1159–1202.

Thaler, R.H., and C.R. Sunstein. 2008. *Nudge*. New Haven: Yale University Press.

Trout, J. 2005. Paternalism and cognitive bias. *Law and Philosophy* 24(4): 393–434.

Tversky, A., and D. Kahneman. 1983. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review* 90(4): 293–315.

Verweij, M., and M. van den Hoven. 2012. Nudges in public health: paternalism is paramount. *The American Journal of Bioethics* 12(2): 16–17.

White, M.D. 2013. *The manipulation of choice: Ethics and libertarian paternalism*. New York: Palgrave Macmillan.

Whitman, D.G., and M.J. Rizzo. 2015. The problematic welfare standards of behavioral paternalism. *Review of Philosophy and Psychology*. doi: 10.1007/s13164-015-0244-5 .

Wilkinson, T. 2013. Nudging and manipulation. *Political Studies* 61(2): 341–355.

[1] Alternatively, Rebonato (2012) states that nudges should be "easily reversible" rather than easily avoidable. Sunstein (2013) and Mills (2015, this issue) argue in favor of the condition of avoidability, rather than reversibility.

[2] For an investigation of the links between freedom of choice and positive and negative liberty, and an analysis of the liberal credentials of libertarian paternalism, see Grüne-Yanoff (2012), Goodwin (2012) and Mills (2013).

[3] For example, Hausman and Welch (2010) argue that "providing information and giving advice treats individuals as fully competent decision maker", and is thus not paternalistic. Dworkin (2014) considers that an intervention must interfere either with freedom or autonomy to be called paternalistic—but providing information arguably does not interfere with either.

[4] However, the influence of behavioral factors on decisions becoming more and more widely known amongst choice architects, arranging a choice architecture without any underlying intention will become a more and more uncommon alternative.

[5] Which states that "the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others" (Mill 1956)

[6] But see Schnellenbach (2012) for the potential danger of consolidating strong norms through voting, even when it is it not in the material self-interest of a majority of voters.

[7] Wilkinson (2013) defines manipulating as intentionally and successfully influencing someone using methods that pervert choice; see Hansen and Jespersen (2013) for a different account of manipulativeness, based on non-transparency.

[8] see also Hansen and Jespersen (2013) who think that the publicity condition should be completed by epistemic transparency.