GAPP, número 25, marzo de 2021
Sección: ARTÍCULOS
Recibido: 24-09-2020
Modificado: 30-12-2020
Aceptado: 04-01-2021
Publicado: 11-03-2021
DOI: https://doi.org/10.24965/gapp.i25.10865
Páginas: 23-37

What is wrong with nudges? Addressing normative objections to the aims and the means of nudges

Nudges, consideraciones normativas de sus objetivos y métodos

Júlia de Quintana Medina
Universitat Autònoma de Barcelona (España)
ORCID: https://orcid.org/0000-0002-4358-3109
juliadequintana@gmail.com

NOTA BIOGRÁFICA

Júlia de Quintana pursues a PhD on the acceptability of nudges in public policy at Universitat Autònoma de Barcelona. She obtained a bachelor's degree in Sociology, awarded with the Special Prize by the Academic Affairs Committee (2014) and a master's degree in Social Policy, Employment and Welfare (2015) from the Universitat Autònoma de Barcelona. Her main research interests are behavioural public policy, institutional design and nudges. Her work covers normative an empirical discussion on decision-making and rationality and social policy and explores the potentialities of behavioural insights in areas such as tax compliance and income guarantee schemes.

ABSTRACT

This paper addresses objections against the aims of nudges–the objectives and legitimacy of nudges–and the means of nudges–the type of influence of nudges. It explores some of the limitations of the main normative criticism in both dimensions and contends that there is nothing inherently wrong with nudges. The paper argues that normative conclusions about their use should discuss nudges ethical implications beyond libertarian paternalism, explore how different nudges affect behaviour and discuss their possible normative drawback surpassing ideal notions of rationality and decision-making.

KEYWORDS

Nudges; libertarian paternalism; rationality; ethics; manipulation.

RESUMEN

El presente artículo discute críticas normativas al uso de *nudges* en política pública, considerando objeciones a sus objetivos y objeciones a su funcionamiento. El articulo explora las limitaciones de ambas objeciones y sostiene que no hay nada intrínsecamente problemático en el uso de *nudges*. El artículo argumenta que las conclusiones normativas sobre su uso deberían discutir sus implicaciones normativas más allá del paternalismo libertario, explorar cómo los diferentes *nudges* afectan el comportamiento y discutir sus implicaciones éticas superando nociones ideales de racionalidad y autonomía.

PALABRAS CLAVE

Nudges; paternalismo libertario; racionalidad; manipulación.

SUMMARY

INTRODUCTION

Nudges are interventions in the decision-making environment that steer individuals towards a particular direction without altering their economic incentives and without forbidding any options. Their implementation is based on evidence from behavioural economics that shows that decision-making is systematically affected by heuristics and cognitive biases; agents' have non-consistent preferences that vary according to the context of choice, and their behaviour is sensitive to social influence (Thaler and Sunstein, 2008). Although their application is gaining increasing presence and importance in several domains, their use remains a controversial topic.

Part of the debate regarding nudges deals with ethical concerns. In the discussion about their ethical acceptability, nudges have sparked accusations of covert paternalism, manipulation, lack of transparency, and overall concerns about having an undermining effect on individuals' freedom of choice and autonomy (e.g. Bovens, 2009; Hausman and Welch, 2010; White, 2013). The ethical debate is so prominent that it seems to indicate that something is particularly wrong with nudges.

Initially, nudges were tied with *libertarian paternalism* (Thaler and Sunstein, 2008). As a normative framework, libertarian paternalism prescribes that nudges should be used to promote specific values and ends. The paternalistic side of the approach aims to change individuals' behaviour to 'make choosers better off, as judged by themselves' (Thaler and Sunstein, 2008, p. 5). The libertarian side of the approach wants to do so without restricting the original set of choices, thereby respecting individuals' freedom of choice. Thaler and Sunstein (2008) remark that these two aspects make nudges a legitimate policy intervention tool. While nudges are now being used beyond this framework, nudges proponents still argue that affecting choices through nudges is unproblematic because nudges steer people towards directions they agree with and do so while maintaining all the options and preserving individuals' ability to opt-out of the nudge.

Despite nudges overall positive intention, it has become apparent in the discussion that these standards are unsatisfactory to support nudges moral acceptability. For the most part, critics find nudges problematic in many grounds and indicate that nudges fail to comply with the original normative standards and carry added normative costs.

Ethical objections to nudges are grounded in many arguments. While arguments are diverse, typically they can be classified as (a) arguments against the aims of nudges, and (b) arguments against the means of nudges. By "the means of nudges" I mean the ways in which they steer people's choices. Objections to the aims of nudges emphasise the difficulties in identifying people's preferences and reorganising individuals' context according to what they want. Objections to the means of nudges usually present nudging as a manipulative strategy that lacks transparency and undermines individuals' autonomy. Although this literature has established essential points to remain vigilant about how nudges are used in public policy, further arguments need to be considered.

This paper explores some of the main normative objections regarding nudges by considering both nudges' objectives and how nudges work. The paper aims to map frequent objections on both dimensions and address their implications and limitations to draw normative conclusions about nudges. In what follows, it is argued that nudges are not intrinsically problematic, neither because of their aims nor because of their means. This paper suggests that normative considerations about nudges require discussing their ethical implications beyond the libertarian paternalism framework, exploring how different nudges affect behaviour and surpassing ideal notions of decision-making and rationality.

The paper proceeds as follows. Section 1 addresses the debate about the aims of nudges. It provides an overview of the main arguments against nudges and responds to the principal objections. Section 2 addresses the objections about the means of nudges, especially the claim that nudges undermine individual autonomy. The discussion identifies two main objections in this direction: *the rationality objection* and *the reflection objection* and discusses their problems to support that nudges are ethically problematic. Finally, the paper ends with some concluding remarks.

1. OBJECTIONS TO THE AIMS OF NUDGES

The first argument in defence of the ethical acceptability of nudges is that nudges improve people's subjective wellbeing. Thaler and Sunstein argue that nudges attempt to influence choices to 'make the choosers better off as judged by themselves' (Thaler & Sunstein, 2008, p. 5). Drawing on findings from behavioural economics, the authors argue that nudges improve individual wellbeing and promote choices that agents would choose if their decisions were not affected by cognitive biases, poor self-control and limited time, information and cognitive abilities. Given these factors, Thaler and Sunstein maintain that policymakers ought to interfere in people's choices and steer them towards the options that their rational self would have chosen free from the influence of decision biases. The argument is vital to defending the ethical value of nudges; the implementation of nudges is deemed ethically legitimate because nudges promote wellbeing, and people agree with the direction in which they are being nudged.

1.1. Nudges and wellbeing, as judged by individuals' themselves?

Thaler and Sunstein claim that nudges steer individual wellbeing as judged by the individuals themselves. However, the general discussion around nudges indicates that this first justification is problematic. Critics argue that to support that nudges promote people wellbeing, Thaler and Sunstein make unfounded assumptions about people preferences and ignore evidence that points otherwise (Gigerenzer, 2015; Infante *et al.*, 2016; Rebonato, 2013; Sugden, 2017; White, 2013). Likewise, critics argue that their use under this justification is problematic for several reasons.

Infante *et al.* (2016) and Sugden (2017) argue that Thaler and Sunstein lack subjective information about what people want, and adequate criteria to prove that people's decisions systematically fail to reflect their true preferences. According to Infante *et al.* (2016), normative behavioural economics uses a preference purification model that assumes that 'an inner rational agent is trapped inside a psychological shell' (Infante *et al.*, 2016, p. 2). When choices are inconsistent with what is expected in rational choice theory, the assumption is that agents have made a mistake due to decision biases. In this context, policymakers need to change people's choices and reconstruct their true preferences. However, the model does not delve into the psychology of choices and only works on the assumption that people's choices do not reflect their preferences.

According to Thaler and Sunstein, it is evident that people prefer to be fit and healthy than overweight and unhealthy, and it is evident that people prefer to cut senseless expenses in the interests of a wealthier future. However, Thaler and Sunstein fail to provide evidence that these inferred preferences reflect people's true preferences. Rational preferences are assumed but are not empirically proven (Sugden, 2017, p. 117).

Sugden (2017) argues that Thaler and Sunstein do not provide satisfactory criteria to distinguish choices resulting from a cognitive bias from decisions motivated by alternative factors. To establish whether a decision is good or bad, Thaler and Sunstein only take its outcome into consideration: if the choice maximises utility, it is rational; if it does not, it is a mistake resulting from a cognitive bias. However, their approach fails to provide appropriate evidence that the choice is an error of judgment. In this regard, Thaler and Sunstein lack empirical evidence about people's true preferences and do not have criteria to identify which choices are affected by cognitive biases.

Similarly, Gigerenzer (2015) also questions the evidence underpinning the fact that nudges promote people wellbeing. Gigerenzer (2015, 2018) argues that behavioural economics lacks empirical evidence to support the notion that people predictably and systematically lack rationality and remarks that decision-making mistakes due to cognitive biases are not as prevalent as Thaler and Sunstein assume. Accordingly, he questions the use of nudges under the justification that they promote people's true preferences. Gigerenzer claims that embracing the standard justification of nudges blames people for their own mistakes while failing to recognise that, in general, their decisions are correct. In this sense, using nudges under the standard justification can lead policymakers to refrain from considering alternative strategies for behavioural change. Gigerenzer argues that, beyond nudges, behavioural evidence supports the use of interventions that intend to educate people, improve their deliberative capacities or correct external factors that may be affecting their decisions (Gigerenzer, 2015, 2018).

In a very confrontational critique, White (2013) also argues that behavioural economics has essential limitations in supporting the claim that nudges promote individual subjective wellbeing because people have different understandings of what is good for them. Given these limitations, White argues that nudges should be rejected because, under the original justification, nudges entail a value substitution that replaces agents'

judgment of what is best for them, with the policymaker's interest over what they should be doing. This is ethically problematic because it gives policymakers *carte blanch* to nudge as they please, without actually justifying when and why nudges promote positive goals.

In summary, critics agree that the claim that nudges promote people's wellbeing is supported by assumptions about human behaviour and rationality rather than on substantial empirical evidence about individuals' preferences. These arguments postulate problems for nudges ethical acceptability. Nudges appear problematic because policymakers might use this claim to conceal an excessive paternalistic agenda or to promote illicit ends.

1.2. Nudges, alternative aims and justifications

When discussing the problems nudges have with fulfilling their original goal, many authors have explored nudges' potential to promote different aims and how their use is ethically acceptable under diverse normative justifications.

Carrying on with Thaler and Sunstein's original intention, many authors have discussed nudge's potential to improve people's wellbeing, regardless of their preferences. Often there can be inconsistencies between what people want and what is right for them. In these situations, it is common for governments to interfere in people's choices and justify this interference under paternalism. Paternalism can be defined, as 'the interference by some outside agent in a person's freedom for the latter's good' (Le Grand & New, 2015, p. 7). This includes cases in which governments interfere in people's choices to promote particular outcomes, even at the expense of agents' preferences. For instance, governments require people to wear seatbelts in cars and helmets when riding motorcycles, prevent people from swimming or skiing in potentially harmful weather conditions and ban drugs to prevent their adverse effects. In line with these interventions, nudges which seek to make people engage in healthy habits or save up money for the future, for instance, seem to follow a paternalistic intention. Indeed, most of the examples illustrated by Thaler and Sunstein seek to help people to get healthier and wealthier by assuming that this is what they want; therefore, policymakers could defend their use under a paternalistic justification and argue that 'nudged individuals are always better off independently of their preferences' (Guala & Mittone, 2015, p. 386)

Beyond individual wellbeing, some authors have considered the advantages of using nudges to tackle social problems. Nudges can be useful in resolving issues where the need for intervention is already justified by traditional economic grounds, for instance, in cases of externalities, public goods and information asymmetry (Chetty, 2015; Loewenstein & Chater, 2017). For instance, Guala & Mittone (2015) argue in favour of using nudges to solve public policy problems, particularly to correct externalities. The authors highlight that Thaler and Sunstein had already considered this option themselves in '*Nudge'* when they mention the potential of nudges to promote energy conservation, organ donation and tax compliance.

Similarly, Loewenstein & Chater (2017) point out that incorporating insight on nudges into issues of public policy can be highly beneficial when it comes to rethinking the tools available for changing behaviour in situations of coordination or social dilemmas. For instance, on matters concerning public goods, the traditional economic framework asserts that it is rational for self-interested individuals to act as free-riders. Drawing on these assumptions, interventions aimed at correcting this behaviour rely exclusively on incentives, bans and the privatisation of public goods (Guala & Mittone, 2015). Behavioural and social sciences have shown that there could be biased motivations underlying what economics classifies as free-rider behaviour. For instance, cognitive biases could be why people fail to cooperate, due to inertia, social influence or the effect of elements of the choice architecture. If this is the case, a nudge could be very useful in improving cooperation and reducing externalities. On a similar note, Nagatsu (2015) introduces the concept of *social nudges* and evaluates the potential of nudges to improve the provision of public goods. In relation to public goods, extensive evidence in social science suggests that individuals are conditional co-operators rather than free-riders, meaning that they have preferences for cooperating if others also cooperate (Bicchieri, 2006; Elster, 2007). In this case, there is no bias; however, using a nudge that communicates the behaviour of others may activate a social norm and could be helpful when it comes to increasing cooperation. These examples illustrate that nudges have the potential to pursue different policy goals and can complement traditional policy interventions.

Using nudges to tackle any policy problem opens up the possibilities of how to legitimise their use. Commenting on alternative ways to justify the implementation of nudges, Kelly (2013, p. 213) maintains that the same techniques that Thaler and Sunstein suggest using to make individuals better off can also be utilised to promote utilitarian and Rawlsian goals. The use of nudges may undermine individual freedom of choice

and may involve paternalism; however, their use may still be legitimate if they improve overall social welfare, reduce inequalities or ensure access to primary goods.

In '*Nudge'*, Thaler and Sunstein outline the rationale for intervening with nudges (the existence of failures of rationality) and define the ethical legitimisation of nudges with libertarian paternalism (promoting individual subjective wellbeing). Section 1.1 has argued that the existing evidence falls short of supporting nudges original argument for ethical acceptability. If people have to agree with their aims for nudges to be legitimated, their implementation is much more restricted than Thaler and Sunstein defend. Likewise, if policymakers use nudges under this justification, its use is problematic and raises relevant normative objections.

However, we should not reject nudges because they fail to promote people's subjective wellbeing. When discussing the aims of nudges, inevitably they are presumed to be linked to libertarian paternalism. In fact, both terms are often confused and used interchangeably (Gigerenzer, 2015; Hansen & Jespersen, 2013; Schubert, 2015). However, nudges and libertarian paternalism have distinct and separable meanings. Libertarian paternalism is a normative framework that aims to promote specific values and ends; nudges are tools to influence behaviour and encourage behavioural change. Nudges should be understood as a policy tool, a tool with practical value for promoting pro-self and pro-social goals, the implementation of which can respond to different normative justifications.

White (2013, p. 83) argues that nudging is not about helping people make better choices, but about getting people to make the choices that 'policymakers want them to make'. The point is that this is not necessarily a problem, particularly if nudges promote ethically consistent goals. It is clear that resolving current public policy problems requires a change in people's behaviour, and that traditional government tools sometimes fail to tackle such issues. Nudges working complementarily with other tools can shape individual behaviour to match different objectives and motivate behavioural change. As far as their aims are concerned, nudges are not inherently problematic.

2. OBJECTIONS TO THE MEANS OF NUDGES

The second argument in defence of nudges' ethical acceptability is the claim that nudges respect agents' freedom of choice. Thaler and Sunstein argue that, in general, people should be free to choose what they want to do according to their preferences. For this reason, libertarian paternalism attempts to design policies 'that maintain or increase freedom of choice' (Thaler & Sunstein, 2008, p. 5).

Nonetheless, several authors note that the fact that nudges do not enforce explicit barriers to liberty is insufficient to address their ethical implications. The majority of critics agree that even though nudges do not block choices, do not modify economic incentives and typically maintain all the available options in a context of choice, they interfere in people's decision-making and diminish their ability to make their own choices. Many found nudges problematic due to how they steer people's behaviour; specifically, they express concerns about nudges' effect on individual autonomy (Hansen & Jespersen, 2013; Hausman & Welch, 2010; Kelly, 2013).

The debate on nudges and autonomy is tricky because it includes different and sometimes contradictory understanding of autonomy as well as normative and descriptive ideas about how individual preferences are formed and should be formed, which factors affect and should affect decision-making, and which elements promote or undermine personal autonomy. Despite the many worries and different conceptualisations of autonomy and decision-making discussed within nudge literature, objections that find nudges problematic due to how they steer choices tend to focus on two ideas:

a) Nudges are problematic because they trigger non-rational psychological mechanisms.
b) Nudges are problematic because they impede or obstruct reflection.

Some critics note that nudges threaten autonomy because they work via irrational mechanisms and take advantage of people's cognitive flaws (Bovens, 2009; Conly, 2012; Hausman, 2018; Hausman & Welch, 2010; White, 2013). Many authors indicate that working through non-rational mechanisms compromises people's autonomy. Other authors suggest that nudges typically work covertly without decision-makers being aware of the nudge and by bypassing or obscuring deliberation (Bovens, 2009; Grüne-Yanoff & Hertwig, 2016), leading to these factors undermining reflectiveness and compromising people's autonomy. In this first theme, the focus is on rational deliberation, and the fear is that nudges trigger non-rational psychological mechanisms. I use the expression "the rationality objection" to represent these worries and objections. The second theme focuses on reflection and conscious deliberation, and nudges appear problematic because

they impede or obstruct it. I use the expression "the reflection objection" to refer to these objections. There is a significant overlap between the arguments; however, differentiating the cases seems to be the best way to address which elements of nudges are problematic.

Below, I discuss these two objections. I analyse why nudges appear problematic, and under which arguments and assumptions those objections hold true. Firstly, I note that in both cases, critics tend to misrepresent nudges. Critics often group nudge interventions into the same category, relying on elements that are difficult to conceptualise and not shared by all nudges. As a result, they assess them as a general category and fail to consider the differences between nudges. Secondly, I argue that idealistic understandings of decision-making and autonomy underpin both objections. Critics employ conceptions of rationality and reflection that lack psychological insights and rely on assumptions unsupported by the empirical evidence on decision-making.

2.1. The rationality objection

The rationality objection states that nudges are problematic because they trigger or take advantage of irrational psychological mechanisms. Hausman and Welch (2010, p. 130) argue that nudges play on 'flaws in human judgment and decision-making to shape people's choices'. Similarly, Bovens (2009, p. 209) argues that what distinguishes nudges from other types of influences in shaping choices is the fact that 'some pattern of irrationality is being exploited'. On the same note, Grüne-Yanoff and Hertwig (2016, p. 153) state that 'what is genuinely novel about the nudging approach […] is the idea of exploiting people's cognitive and motivational deficiencies'. Conly (2012, p. 30) also remarks that, when nudging, 'rather than regarding people as generally capable of making good choices, we outmanoeuvre them by appealing to their irrationality, just in more fruitful ways'.

In general, authors agree with the fact that influencing choices by triggering irrational responses is ethically problematic. According to Hausman and Welch (2010, p. 128), when shaping choices 'does not take the form of rational persuasion, autonomy–the extent to which individuals have control over their own evaluation and deliberation–is diminished'. On a similar note, Bovens (2009, p. 209) points out that 'there is something less than fully autonomous about the patterns of decision-making that *Nudge* taps into' and argues that 'when we are subject to the mechanisms that are studied in "the science of choice", then we are not fully in control of our actions'.

The above citations suggest that the problem with nudges is the fact that they exploit irrationality. For example, framing devices that highlight losses appear harmful because they affect behaviour by exploiting loss aversion and, therefore, change behaviour by exploiting an irrational bias without involving rational reasons. Similarly, default rules appear to be problematic because they do not constitute a rational reason to change preferences over options (Bovens, 2009). Likewise, the communication of social norms works because it triggers an irrational response and is problematic because agents do not make choices based on a rational consideration (Bovens, 2009; Hausman & Welc, 2010; Schubert, 2015). In the rationality objection, the essential idea is that exploiting irrationality compromises people's autonomy.

The objection to nudges on the basis that they exploit irrationality suffers from two main problems. Firstly, not all nudges exploit irrationality and stipulating which interventions do and which do not appears to be somewhat challenging. Secondly, to argue that nudges exploit irrationality and exploiting irrationality compromises people autonomy critics rely on rational choice theory as the normative foundation of behaviour, a framework with notorious problems.

2.1.1. Nudges and rationality

Let us start by considering the relationship between nudges and rationality. The most common answer for defending nudges against the rationality objection is to argue that not all nudges exploit irrationality. Sunstein (2015b), for instance, responds to critics by claiming that many nudges, such as, education campaigns, informational campaigns, reminders, warnings and the provision of feedback are interventions that engage rational deliberation and do not constitute a threat to people's rationality[1]. Sunstein employs a broad

[1] There is a problem of conceptualisation within this debate. It is difficult to articulate whether these interventions should count as nudges because there are no precise definitions of a nudge. The border between which interventions are nudges and which interventions are not is diffuse. Several interventions count as nudges because they do not change economic incentives and do not ban or exclude any option; however, it is unclear how they relate with rationality and decision-making biases and whether they should be considered nudges.

definition, which fits many different types of interventions that do respect rational deliberation. If one adopts his definition, then indeed many nudges respect rationality. However, even if we adopt Sunstein's approach, the fact that some so-called "nudges" can be considered generally unproblematic does not resolve concerns about those interventions that work by exploiting cognitive flaws. In recent publications, critics seem to direct the rationality objection only at those interventions that rely on people's cognitive biases. For instance, Hausman (2018, p. 18) uses the term "nudging" to reference 'changing the choice circumstance to neutralise or to exploit deliberative foibles' and explicitly distinguishes nudges from other ways of steering choices such as information, education or deceiving. However, even when more refined, it is challenging to articulate which decision-making factors neutralise or exploit irrational biases.

In '*Nudge'* Thaler & Sunstein (2008, p. 37) stress the idea that human behaviour is "nudgeable", which means it can be easily influenced through mechanisms not considered within the traditional economics framework. The authors mention different factors and classify them into three groups: *biases and blunders*, *temptation*, and *following the herd* (Mongin & Cozic, 2018; Thaler & Sunstein, 2008).

Broadly, by *biases and blunders* they refer to the influence of heuristics such as representativeness, anchoring and adjustment, and availability; and biases such as overconfidence, loss aversion, and framing effects (Tversky & Kahneman, 1974; Kahneman & Tversky, 1979). By *temptation*, they refer to a lack of will power and general failures to control impulses and maintain self-control. By *following the herd*, they emphasise the significance of social influence and social norms in shaping agents' decision-making. Within behavioural economics, these factors are conceived as deviations from rational choice and regarded as non-rational. As a result, designing choice architecture in ways that engage with these psychological factors implies that people's choices are affected by irrational factors. However, whether these factors are irrational is not a straightforward point. The factors described in '*Nudge'* as decision-making flaws do not hold true if interpreted according to alternative understandings of rationality (Gigerenzer, 2015; Schubert, 2015; Mongin & Cozic, 2018).

Firstly, the *biases and blunders* category refers to factors such as framing effects, loss aversion and other psychological biases found under the heuristics and biases research programme. Within behavioural economics, framing effects and loss aversion are presented as typical examples of decision-making biases. They occur when an agent's preferences between two identical sets of alternatives vary depending on how options are described, particularly if they are described as losses rather than gains. These findings violate the axioms of the *expected utility model* and are accordingly classified as biases. However, adopting an alternative approach, Gigerenzer (2008) maintains that, in cases where two definitions of the same situation are framed, they are logically equivalent but not informationally equal. Given contextual and cognitive constraints, agents use *fast and frugal heuristics* for breaking down pertinent information and decide between options. In these situations, decisions shaped by framings effects are not irrational but *ecologically rational*, i.e., rational in a particular context. The fast and frugal research programme generally challenges the findings of the heuristics and biases research programme. Research on *ecological rationality* establishes that many supposed irrational biases are "ecologically rational" across environments, given conditions of uncertainty and limited cognitive resources.

The *temptation* category refers to issues of self-control, failures to resist temptation and time inconsistency. According to different conceptions of rationality, these factors often qualify as failures of rationality. Decision-making research on temporal choice has verified a systematic preference for small but imminent rewards over greater but delayed rewards. Agents discount future utility, a phenomenon labelled as *temporal discounting.* In behavioural economics, research on temporal choices follows the *hyperbolic discounting model*, a model that asserts that when agents face a choice between an inferior early option and a superior later option, they prefer the latter when both options are remote in time but switch to preferring the former as the time for both options approaches. The change of preferences appears irrational because agents' preferences are not consistent over time (Laibson, 1997). However, rational factors could also explain choice inconsistency and preference reversal. Inconsistent changes in preferences over time can result from imperfect foresight, lack of information, or recently learned information, factors that might induce individuals to change or update their preferences. Therefore, while inconsistent time preferences often reflect irrationality, it is empirically difficult to identify in which situations this is the case and to distinguish these cases from rational cases of preference reversal.

In terms of *following the herd*, decision-making research provides extensive evidence as to why following a social norm or following the behaviour of others is unlikely to be irrational. According to broader notions of rational choice theory that question the assumption of unbounded self-interest, game theory research

suggests that as far as public goods are concerned, individuals have conditional preferences for conformity meaning that they prefer cooperation if a sufficient number of others also cooperate (Bicchieri, 2006; Elster, 2007). Nagatsu (2015) maintains that nudges that provide information about the behaviour of the majority mobilise these preferences and trigger a rational response. Gigerenzer (2015) also contends that observing and following other's behaviour in the context of uncertainty and limited information is following a social heuristic, which is an "ecologically rational" strategy. Likewise, Hedström (2006) argues that observing and copying other behaviours in cases of limited information is a rational strategy, a mechanism usually called "rational imitation".

The examples described highlight alternative interpretations of the same phenomena as rational or irrational, depending on the understanding of rationality employed[2]. Some nudges are designed to engage with psychological mechanisms traditionally regarded as non-rational within economics. As a result, conceptualising nudges as interventions that exploit irrationality requires accepting the rational choice model as the normative model of behaviour and distinguishing good (rational) influences from bad (non-rational) ones. However, there is considerable disagreement about identifying rational and non-rational factors and controversy about which nudges trigger rational or non-rational mechanisms. The objection to rationality already excludes nudges that most resemble rational persuasion, such as reminders, warnings, some types of information and some types of educational campaigns. However, doubts about how nudges relate to rationality are also relevant for interventions that exploit cognitive weaknesses and are typically classified as nudges.

2.1.2. Rationality and autonomy

Let us now consider why exploiting irrational factors is problematic in terms of autonomy. Critics identify nudges as negative influences because they do not engage rationality. Rationality appears to be the crucial element of autonomous choice within this approach; individuals are expected to engage in rational thinking, process all the relevant information and act according to their consistent preferences. However, empirical evidence on decision-making indicates that this approach relies on a somewhat idealistic and heroic view of rationality and decision-making, which does not match reality.

Rational choice theory is a *substantive* theory of rationality (Simon, 1997). It uses a specific set of axioms to study and model behaviour and employs assumptions about agents' internal decision-making processes, namely perfect rationality, perfect foresight, consistent preferences and unlimited computation abilities. The expected utility model works with these axioms and explains and predicts behaviour *as if* subjects behave accordingly. Rational choice theory serves as a theory for modelling behaviour but does not account for the actual process of decision-making. It does not delve deeper into the decision-making process and does not explore the truthfulness of axioms.

While a substantive theory of rationality can be useful in decision-making research, particularly when constructing models and predicting behaviour, when the same conception is used to inform the ethical debate on autonomy, it is much more problematical. The rational choice theory model faces significant criticism both in descriptive and normative terms. Extensive experimental findings question its descriptive power (e.g. Angner & Loewenstein, 2007; DellaVigna, 2009), and the framework attracts significant criticism when used as a normative theory of decision-making (Elqayam & Evans, 2011; Gigerenzer, 2015). A conception of autonomy as rationality based on a model that does not accurately describe agents' decision-making and is notorious for its detractors is not suitable for judging nudges' moral acceptability.

Those committed to the standard model of rationality may still argue that nudges are problematic. They may argue that, ideally, people should be rationally persuaded, and that governments and policy institutions should prioritise the use of interventions that influence behaviour via rational reasons such as education, information, monetary incentives and coercive measures. However, committing to rational choice has some limitations.

Firstly, non-rational factors will always affect behaviour. Most actions are affected by a multiplicity of motives; some of these motives qualify as rational, others not. We cannot expect people to only act for rational reasons; non-rational reasons play an equal part in and are relevant to decision-making. Even in the absence

[2] In some cases, it is not even a different concept of rationality, but using rational choice theory, with some informational assumptions, to explain the phenomenon. For example, following the behaviour of others can be explained with the rational imitation mechanism, in which, in contexts of limited information, what others are doing signals the rational (or utility maximation) course of action (Hedström, 2006).

of nudges, non-rational factors usually affect choices, so it is unclear why these factors should be considered unsuitable for use in changing decisions. Secondly, nudges do not intend to replace other strategies, and their use is compatible with trying to change people behaviour by using rational strategies. Likewise, in some cases, nudges may have some comparative advantages over other alternatives. For instance, education campaigns have limitations to tackle some issues, are more expensive and do not always induce behavioural change in the short term (Conly, 2012; Datta and Mullainathan, 2014) The use of more coercive strategies such as regulations and sanctions can backfire due to crowding out non-egocentric deliberation and intrinsically motivated compliance (Pettit, 1996). Similarly, the provision of monetary incentives provides extrinsic motivations for doing certain behaviours; and in some contexts, monetary incentives tend to discourage those that were naturally drawn to comply and frame a situation as a monetary issue which legitimises free-rider behaviour and can reduce the chances of peer punishment for non-compliance (Gneezy *et al.*, 2011). In these cases, using nudges might be beneficial to surpass these limitations.

The rationality objection essentially argues that nudges do not respect rational decision-making and undermine people's autonomy. These objections only hold true under assumptions on behaviour based on the rational choice model. According to the model, individuals ought to behave and actually do behave rationally; thus, nudges that exploit irrationality compromise their autonomy. However, when one acknowledges the limitations of rational choice theory and considers alternative understandings of rationality, the rationality objection loses strength (Schmidt, 2019). Firstly, arguing that nudges trigger irrational mechanisms is not a clear-cut empirical distinction but a normative one that requires a normative assessment of decision-making factors according to the rational choice theory framework. Secondly, the account of autonomy as rationality is too idealistic and normatively charged, and empirical research emphasises that it is based on unrealistic assumptions (Felsen & Reiner, 2015; Mills, 2015; Schubert, 2015; Schmidt, 2019).

2.2. The reflection objection

The reflection objection states that nudges are harmful and compromise individuals' autonomy because they actively reduce agents' engagement in the decision-making process. Reflection means that agents are aware of the factors that influence their choices, engage with these factors and exert some sort of effort and deliberation into decision-making. Some authors stress that nudges are unnoticeable, work unconsciously and steer people towards a specific choice with limited consciousness and limited effort. As a result, critics argue that nudges compromise people's reflectiveness.

Blumenthal-Barby and Burroughs (2012), state that nudges can imply manipulation when they are difficult to perceive:

> 'Manipulation occurs when one influences another by bypassing their capacity for reason, either by exploiting nonrational elements of psychological makeup or by influencing choices in a way that is not obvious to the subject' (p. 5)

Bovens (2009) notes that nudges tend to be hard to perceived and work best when unnoticeable:

> 'The psychological mechanisms that are exploited in Cafeteria and in Save More tomorrow[3] typically work better in the dark. If we tell students that the order of the food in the Cafeteria is rearranged for dietary purposes, then the intervention may be less successful' (p. 3)

Grüne-Yanoff and Hertwig (2016) state that nudges can produce an automatic response without agents' active attention:

> 'in a nudge, the policymaker does not rely on the agent's ability to stop or override a targeted behavior or cognition. Instead, the nudge intervention can 'remedy' an individual's actions without the individual making an active contribution. It is this feature that leads critics to argue that nudges are manipulative and that they violate autonomy and dignity' (p. 176)

Baldwin (2014) states that some nudges compromise autonomy because they work without awareness and reflection:

[3] The cafeteria nudge references the reorganisation of the food chain in a Cafeteria in a way that makes healthy options easier and more convenient to select. The Save More tomorrow references a programme that increases pension contributions with each pay raise by default (Thaler & Sunstein, 2008).

> 'the targeted individual's behavioural or volitional limitations and 'automatic' responses will in practice lead him or her to accept the nudge with limited awareness and reflection.'(p. 836)

These quotes express worries about nudges impeding reflectiveness[4]. In contrast to the examples discussed in the last section, it is easy to appreciate the difference with worries on rationality. For instance, nudging by using framing effects is an intrusion on autonomy because agents cannot identify that there is an intervention aimed at shaping their behaviour and fail to recognise how the frame affected their decision. Baldwin (2014) argues that framing devices can shape behaviour 'in a manner that is resistant to unpacking in so far as assessing the nature and extent of the nudge is not readily achieved by refection' (p. 836). In this case, the worry is not about rationality but reflectiveness. Similarly, defaults appear problematic because they can work without individuals perceiving the intervention or its effect. Therefore, defaults undermine autonomy, not because they exploit a cognitive bias, but due to their potential to influence choices without awareness and reflection (Smith, Goldstein, and Johnson 2013). By contrast, the communication of social norms is a nudge that does not appear as problematic because it engages reflection: agents can identify the nudge, and it works through a process that involves some degree of reflection (Bovens, 2009; Hansen & Jespersen, 2013). As a result, even though communicating social norms might activate a non-rational response, it is a less problematic nudge because it engages reflectiveness.

The key aspects of concern regarding the reflection objection are: (a) nudges bypass deliberation and affect behaviour unconsciously, and (b) nudges lack transparency and have to operate unnoticeably to work. These two features raise concerns about nudges pushing people into doing things without reflectiveness, thereby undermining their autonomy. However, the reflection objection has two main problems, just as the rationality objection does. Firstly, it is unclear which nudges bypass deliberation, and are unnoticeable. Secondly, it considers reflectiveness to be a crucial component of autonomous choice, a notion which appears to be too idealistic when the psychological evidence is taken into account. I shall start by commenting on the first issue, and then return to the discussion on reflectiveness, decision-making and autonomy.

2.2.1. Diversity of nudges and ethical implications

With regard to nudges impeding people's reflectiveness, there appears to be a consensus on the features that make nudges problematic: (a) they bypass deliberation (b) they are not transparent. While both claims are frequent in nudge literature, many authors also acknowledge that not all nudges have these features. Sunstein points this out by emphasising 'the importance of having a sufficiently capacious sense of the category of nudges, and an appreciation of the differences among them' (Sunstein 2015a, p. 513) and uses this argument to defend nudges against objections. In line with his argument, many authors have paid more attention to the fact that conclusions about the moral acceptability of nudges should focus more on each nudge's specific features.

Current literature offers different classifications of nudge interventions. The most well-received and commonly used classification of nudges is the differentiation between System 1 and System 2 nudges. The distinction follows Kahneman's (2011) account of the dual-process cognitive theory, which describes two distinct systems underlying human reasoning: System 1 and System 2. System 1 is fast, automatic, uncontrolled, and unconscious. System 2 is slow, conscious, reflective, and controlled (Kahneman, 2011). When applied

[4] The worries on reflection underpin worries on manipulation, transparency, and dignity. I will not address these objections specifically because I suggest that they boil down to worries about nudges' effect on reflectiveness. For instance, the account of manipulation used in nudge literature always references nudges working unconsciously and 'in the dark', i.e., working without the full attention and awareness of agents. Because the influence operates in the dark, the nudge is non-transparent and implies manipulation. However, manipulation can be understood in different ways and can take different forms. Firstly, manipulation implies that A has to pressure B to do something that B does not want to do. In the case of nudges, while we cannot argue that people always want to be nudged towards an outcome, we cannot say that they oppose it. Likewise, manipulation can happen transparently, by shaping arguments, exploiting emotions and deceiving people. In these forms, manipulation does not operate 'in the dark' and has nothing to do with whether agents notice an intervention's effect or influence (Wilkinson, 2013). In nudge literature, claims of manipulation and transparency relate to the degree of reflectiveness that agents have in decision-making; thus, it is better to address these objections by examining the effects of nudges on reflectiveness. Similarly, objections related to dignity and individual responsibility also reflect concerns about how nudges affect reflectiveness. Detractors of nudges maintain that nudges do not help us to learn to make better choices in the future, because they override our conscious reasoning and block the learning process (White, 2013, p. 102). This leads some to worry that nudging infantilises individuals and undermines their responsibility and dignity (Bovens, 2009; Gigerenzer, 2015; White, 2013). Again, these objections raise concerns about how nudges shape choices, and how much reflection they engage; therefore, to allay such worries, it would be better to address them by discussing the effects of nudges on reflectiveness.

to nudges, the schema distinguishes between System 1 and System 2 nudges. System 1 nudges tend to be described as unconscious, non-deliberative influences that override human agency. System 2 nudges are described as reflective triggering influences that engage agency. The two types are also referred to as non-educational and educational nudges respectively (e.g. Hertwig and Ryall 2016; Sunstein, 2017). Examples of System 1 nudges include the use of defaults and the design of physical options in the decision-making environment. Examples of System 2 nudges include disclosing information or using educational campaigns. System 2 nudges implicitly tend to be seen as good nudges while System 1 nudges appear to be more morally problematic.

The distinction between System 1 and System 2 serves as a good schema to argue against objections that present nudges as interventions that always bypass deliberation, tend to prompt an unconscious or unreflective response and are usually unnoticed. The distinction has been adopted by proponents of nudges and is regularly employed to differentiate between nudge interventions and address their ethical implications. The distinction can be found in research on the empirical performance and effectiveness of nudge interventions (e.g. Hollands et al. 2013; Smith, Goldstein and Johnson 2013) and in reserch on *attitudes towards nudges* to study whether people accept nudges and how they judge different nudge interventions (e.g. Felsen, Castelo, and Reiner, 2013; Hagman et al. 2015; Jung and Mellers 2016; Sunstein, 2017).The classification is still preliminary and requires further research on the psychological mechanisms behind different nudges. However, it is a good starting point to distinguish how different nudges affect behaviour and what this implies in terms of moral implications.

Alternatively, other authors have also presented classifications to discuss the moral implications of different nudges. For instance, Baldwin (2014) develops a taxonomy of nudge interventions that classifies nudges in three degrees: First Degree, Second Degree and Third Degree nudges, depending on how an intervention engages autonomy. Baldwin operationalises autonomy as the way in which nudges take reflective decision-making into account (Baldwin, 2014, p. 835). Similarly, Saghai (2013) distinguishes between interventions that entirely or partly bypassed deliberation. He argues that it is often assumed that nudges trigger an automatic cognitive process, however, this is not always the case. On a similar note, Hansen & Jespersen (2013) present an epistemic distinction of nudges according to how different nudges engage deliberation and whether the interventions are transparent. With regard to deliberation, they classify nudges as Type 1 and Type 2. The aim of Type 1 nudges is to change behaviour without involving reflective thinking, while Type 2 nudges involve reflective thinking, and mainly influence behaviour resulting from some degree of deliberation. With regard to transparency, they define a transparent nudge as a 'nudge provided in such a way that the intention behind it, as well as the means by which behavioural change is pursued, could reasonably be expected to be transparent to the agent being nudged' (Hansen & Jespersen, 2013, p. 17). In their classification, both Type 1 and Type 2, nudges can be transparent and un-transparent. In their framework, Type 1– non-transparent are seen as more ethically problematic. Finally, Bovens (2009) also explores the differences between nudges in terms of transparency and suggest that nudges that are not directly evident to the subjects are more morally problematic.

The available classifications differ in the criteria they use and how they classify nudge interventions. However, they all agree that nudges operate differently and that considering these differences is crucial to assess their ethical acceptability. Further research on how nudges affect behaviour is needed to identify those nudges that impede reflectiveness and can be ethically problematic.

2.2.2. Reflectiveness and autonomy

The reflection objection establishes that nudges are harmful because they impede reflection by bypassing or obstructing deliberation and working in a manner that is not transparent and cannot be identified by subjects. Critics emphasise that agents should be conscious of the factors that affect their choices and engage in some reflection to produce a choice and therefore contend that nudges compromise people's autonomy. However, the conception of autonomy as reflectiveness holds high standards about how decision-making ought to be and overlooks issues arising from the context of choice that might affect the decision-making process and undermine people's autonomy. Different responses to the worries present in the reflection objection indicate that focusing only on how nudges affect the decision-making process is limited to assess their ethical implications on autonomy.

Research suggests that the vast majority of our choices are affected by unconscious influences (Felsen & Reiner, 2015; Newell & Shanks, 2014; Uhlmann, Pizarro, & Bloom, 2008). In general, we act upon many factors

that we do not perceive, and we are rarely fully aware of all the factors that affect our decisions (Conly, 2012; Felsen, Castelo, & Reiner, 2013). Therefore, it is unclear why unconscious and covert influences should be labelled as being problematic. At the same time, even if we are unaware of the intervention or relevant psychological mechanisms that bring about a particular choice, we are usually aware of the choices we make and we can reflect on them (Uhlmann *et al*., 2008). So, even if non-conscious and covert factors affect our decision, in a vast majority of contexts 'we still consider ourselves as the authors of these choices, *post hoc*' (Nagatsu, 2015, p. 489).

A conception of autonomy formulated primarily based on the fact that people have to be overtly influenced and aware of the factors that affect their choices is inconsistent with empirical evidence on decision-making (Felsen & Reiner, 2011, 2015). Many factors that affect agents' choices are not directly evident to them and influence their behaviour unconsciously. At the same time, the fact that some nudges might influence behaviour covertly is far from meaning that agents will be completely unaware of their resulting behaviour or choice. They may be unaware of the nudge, or the mechanism, but they are unlikely to be unaware of their own behaviour.

Several authors note that, when assessing nudges' ethical value, critics make implicit assumptions about how agents' reasons and motivations come about, particularly about how internal and external factors influence their choices. The requirement that agents have to be aware of the factors that affect their decisions is associated with the idea that there are "pure" reasons that guide people's decisions. Implicitly, critics rely on an ideal notion that assumes that there are "authentic" internal causes for action that should be respected and preserved to guide individual decisions (Felsen & Reiner, 2015; Fischer & Lotz, 2014; Schubert, 2015). Detractors of nudges seem to presuppose that internal causes are defined, stable and perfectly distinguishable from external influences; consequently, nudges are external, intentional influences that disturb pure internal reasons and, therefore, undermine people's autonomy.

By contrast, Thaler and Sunstein (2008) argue that choices are always affected by the disposition of options in the context of choice as a fundamental argument in defending nudges' acceptability. The authors claim that the effect that choice architecture has on behaviour is inevitable, in the sense that individuals' decisions will be affected by other arbitrary factors beyond their control in the absence of nudges. Building on this argument, Thaler and Sunstein (2008) emphasise that opposing nudges is nonsensical, because people are always being nudged, by which they mean always being influenced by non-controlled factors in the choice architecture.

Similarly, Schubert (2015, p. 8) argues that, 'when debating the ethics of nudging, we should stop idealising the institutional *status quo*'. People's decisions are highly shaped by their context of choice. In the absence of nudges, subjects' actions are not guided by "pure" and "authentic" reasons but by reasons that reflect the influence of external and internal factors and constraints (Schmidt & Engelen, 2020).

Detractors of nudges point out that the inevitability of choice architecture is not sufficient to grant nudges free licence; a non-intentional, random influence does not have the same implications as a deliberate intervention aimed at steering people's behaviour towards a specific end. Covert influences cannot be applied acritically just because other non-intentional covert influences generally affect behaviour. The intention and rationale of the intervention have to be discussed if an administration is planning on using nudges. However, confirming the fact that many non-conscious and non-transparent factors affect choices relaxes objections to reflectiveness.

People's choices, values and actions are not the consequence of deliberation based on purely internal motives, but rather the result of a combination of internal and external factors. The notion of autonomy of reflectiveness loses significance in light of the evidence that individuals adjust their preferences and aspirations to their possibilities (Elster, 1983), that the context of choice heavily influences people's behaviour and the confirmation that many processes, whether conscious, unconscious, intentional or non-intentional, affect individual desires and choices.

Several philosophical notions of autonomy emphasise that individuals have to be able to make the choices they want to make in order to be autonomous and that these choices should reflect their true preferences and be authentic and consistent with their 'higher-order desires' (Bovens, 2009; Dworkin, 1988; Felsen & Reiner, 2011). The arguments under the reflection objection stress that nudges compromise autonomy because they corrupt the formation of preferences and desires, and people's choices no longer reflect their authentic preferences. However, these objections place too much emphasis on the process of making choices and overlook how much people's choices are shaped by the context and how many factors beyond the agents' control tend to influence their decisions.

Building on the findings that emphasise the effect that the context of choice has on shaping choices, and the fact that agents struggle to control the factors that affect their behaviour, there is growing recognition of the fact that autonomy requires a combination of internal and external conditions (Mills, 2013, 2015). Proponents of nudges defend that nudges can work as external, intentional sources of influence that will ensure a

better relationship with internal influences, leading to an increase in personal autonomy. For instance, nudges that aim to close the intention-action gap (i.e., the gap between what people intend to do and what they do) are nudges that promote people's best interest in line with their second-order desires. Likewise, in cases of choice overload (i.e., situations in which individuals face too many options and struggle to process the information), agents feel overwhelmed and tend to avoid choosing; consequently, a nudge might act as a choice enabler. In contexts in which people lack feedback or in situations in which there is a gap between the causes and consequences of actions, nudges might be helpful in closing this gap. Likewise, in situations involving a high degree of difficulty or complexity, nudges could be used to promote informed choices or facilitate choice.

Conceptions of autonomy that stress that autonomous decisions should be consistent with higher-order desires (e.g., Dworkin, 1988), balance a trade-off between letting people reflect on their own choices and helping them achieve outcomes in line with their higher-order desires. There is tension between respecting people's reflectiveness and promoting positive outcomes that align with their preferences to promote their autonomy. Excessive attention to how nudges affect decision-making might detract from the fact that they can also work to promote autonomy.

In conclusion, the arguments under the reflection objection express valid doubts about how nudges affect the decision-making process and in which ways nudges can compromise people's autonomy. However, nudges cannot generally be rejected on the bases that they undermine reflectiveness. Firstly, nudges differ in how they affect choices, and the assessment of their ethical implications should pay more attention to how different nudges operate. Secondly, the notion of autonomy as reflectiveness is too idealistic. By emphasising that agents need to be aware of nudges, critics implicitly assume that people make decisions based on purely internal reasons. However, given the effect that choice architecture has on people's choices, this does not make conceptual sense. Likewise, the objections on reflection put too much emphasis on how nudges affect the decision-making process and tend to overlook how retaining reflectiveness might be counterproductive to autonomy and agents' overall welfare. Concerns on reflection are important, but they are not sufficient to discard the use of nudges in policymaking.

CONCLUDING REMARKS

The normative implications of nudges have been largely debated in the past decade. Objections to their use have been prominent and have led to various conclusions on their ethical acceptability as tools to change behaviour. This paper has attempted to argue that, while the current literature generally highlights that nudges have an intrinsically problematic normative nature, this is not the case.

As public policy tools, it is a mistake to assume that nudges are only feasible within the libertarian paternalism framework. In that respect, the paper emphasises the need to discuss alternative justifications for their implementation, to think about how nudges can contribute to better public policy performance by complementing other policy tools and promoting already relevant policy goals. Policymakers should be clear about the rationale and intention behind applying nudges, but these can vary and so can their ethical acceptability.

Beyond the legitimacy of their objectives, some claim that nudges are objectionable because of the ways in which they change behaviour infringe autonomy. The paper has grouped the central claims against the means of nudges into two categories: the rationality objection and the reflection objection to address these concerns. In response to the two complaints, the paper defends that objectors tend to rely on a generalise and misrepresented idea of nudges and on unrealistic understandings of autonomy not supported on empirical bases. Revising the nature of nudges and adopting empirical-based notions of decision-making highlights limitations of both objections to sustain that nudges infringe autonomy.

Several questions on the ethics of nudges remain to be answered. The use of nudges in public policy requires a discussion on whether and how to nudge, this is, which aims nudges should promote, how different nudges work and affect decisions-making and in which contexts their use is normatively acceptable. However, this discussion should not start from a misleading conception of nudges as essentially problematic from an ethical point of view.

REFERENCES

Angner, E., & Loewenstein, G. (2007). Behavioral economics. In U. Mäki (Ed.), *Handbook of the philosophy of science: Philosophy of economic*, (pp. 641-690). Elsevier.

Baldwin, R. (2014). From Regulation to Behaviour Change: Giving Nudge the Third Degree. *Modern Law Review, 77*(6), 831-857. https://doi.org/10.1111/1468-2230.12094

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Blumenthal-Barby, J. S., & Burroughs, H. (2012). Seeking Better Health Care Outcomes: The Ethics Of Using The «Nudge». *The American Journal Of Bioethics*, *12*(2), 1-10. https://doi.org/10.1080/15265161.2011.634481

Bovens, L. (2009). The Ethics of Nudge. In T. Grüne-Yanoff, & S. O. Hansson (Eds.), *Preference change. Approaches From Philosophy, Economics And Psychology* (pp. 207-219). Springer. https://doi.org/10.1007/978-90-481-2593-7_10

Chetty, R. (2015). *Behavioral Economics and Public Policy: A Pragmatic Perspective* (NBER Working Paper Series, 20.928). National Bureau of Economic Research. https://doi.org/10.3386/w20928

Conly, S. (2012). *Against Autonomy. Justifying Coercive Paternalism*. Cambridge University Press. https://doi.org/10.1017/CBO9781139176101

Datta, S., & Mullainathan, S. (2014). Behavioral Design: A New Approach to Development Policy. *The Review of Income and Wealth*, *60*(1), 7-35. https://doi.org/10.1111/roiw.12093

DellaVigna, S. (2009). Psychology And Economics: Evidence From The Field. *Journal of Economic literature*, *47*(2), 315-372. https://doi.org/10.1257/jel.47.2.315

Dworkin, G. (1988). *The Theory And Practice Of Autonomy*. Cambridge University Press. https://doi.org/10.1017/CBO9780511625206

Elqayam, S., & Evans, J. S. B. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*(5), 233-248. https://doi.org/10.1017/s0140525x1100001x

Elster, J. (2007). *Explaining Social Behavior. More Nuts and Bolts for the Social Sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9780511806421

Fischer, M., & Lotz, S. (2014). *Is Soft Paternalism Ethically Legitimate?–The Relevance Of Psychological Processes For The Assessment Of Nudge-Based Policies* (CGS Working Paper Series, 5-2). Cologne Graduate School in Management, Economics and Social Sciences. https://cgs.uni-koeln.de/fileadmin/wiso_fak/cgs/pdf/working_paper/cgswp_05-02.pdf

Felsen, G., & Reiner, P. B. (2011). How The Neuroscience of Decision Making Informs Our Conception of Autonomy. *AJOB Neuroscience*, *2*(3), 3-14. https://doi.org/10.1080/21507740.2011.580489

Felsen, G., Castelo, N., & Reiner, P. B. (2013). Decisional enhancement and autonomy: public attitudes towards overt and covert nudges. *Judgment and Decision Making*, *8*(3), 202-213. http://journal.sjdm.org/12/12823/jdm12823.pdf

Felsen, G., & Reiner, P. B. (2015). What can Neuroscience Contribute to the Debate Over Nudging? *Review of Philosophy and Psychology*, *6*(3), 469-479. https://doi.org/10.1007/s13164-015-0240-9

Gigerenzer, G. (2008). *Rationality For Mortals: How People Cope With Uncertainty*. Oxford University Press.

Gigerenzer, G. (2015). On the Supposed Evidence for Libertarian Paternalism. *Review of Philosophy and Psychology*, *6*(3), 361-383. https://doi.org/10.1007/s13164-015-0248-1

Gigerenzer, G. (2018). The Bias Bias In Behavioral Economics. *Review of Behavioral Economics*, *5*(3-4), 303-336. http://dx.doi.org/10.1561/105.00000092

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives*, *25*(4), 191-210. https://doi.org/10.1257/jep.25.4.191

Grüne-Yanoff, T., & Hertwig, R. (2016). Nudge Versus Boost: How Coherent are Policy and Theory? *Minds and Machines*, *26*(1-2), 149-183. https://doi.org/10.1007/s11023-015-9367-9

Guala, F., & Mittone, L. (2015). A Political Justification of Nudging. *Review of Philosophy and Psychology*, *6*(3), 385-395. https://doi.org/10.1007/s13164-015-0241-8

Hagman, W., Andersson, D., Västfjäll, D., & Tinghög, G. (2015). Public Views On Policies Involving Nudges. *Review of philosophy and psychology*, *6*(3), 439-453. https://doi.org/10.1007/s13164-015-0263-2

Hansen, P. G., & Jespersen, A. M. (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*, *4*(1), 3-28. https://doi.org/10.1017/s1867299x00002762

Hausman, D. M. (2018). Nudging And Other Ways Of Steering Choices. *Intereconomics*, *53*(1), 17-20. https://doi.org/10.1007/s10272-018-0713-z

Hausman, D. M., & Welch, B. (2010). Debate: To Nudge or Not to Nudge*. *Journal of Political Philosophy*, *18*(1), 123-136. https://doi.org/10.1111/j.1467-9760.2009.00351.x

Hedström, P. (2006). Explaining Social Change: An Analytical Approach. *Papers. Revista de Sociologia*, 80, 73-95. https://doi.org/10.5565/rev/papers/v80n0.1770

Hertwig, R., & Ryall, M. D. (2016). *Nudge vs. Boost: Agency Dynamics Under 'Libertarian Paternalism'*. Social Science Research Network (SSRN). http://dx.doi.org/10.2139/ssrn.2711166

Hollands, G. J., Shemilt, I., Marteau, T. M., Jebb, S. A., Kelly, M. P., Nakamura, R., Suhrcke, M., & Ogilvie, D. (2013). Altering micro-environments to change population health behaviour: towards an evidence base for choice architecture interventions. *BMC Public Health*, *13*(1), Article 1218. https://doi.org/10.1186/1471-2458-13-1218

Infante, G., Lecouteux, G., & Sugden, R. (2016). Preference purification and the inner rational agent: A critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, *23*(1), 1-25. https://doi.org/10.1080/1350178X.2015.1070527

Jung, J. Y., & Mellers, B. A. (2016). American attitudes toward nudges. *Judgment and Decision Making*, *11*(1), 62-74. http://journal.sjdm.org/15/15824a/jdm15824a.pdf

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263-291. https://doi.org/10.2307/1914185

Kahneman, D. (2011). *Thinking, Fast And Slow*. Macmillan.

Kelly, J. (2013). Libertarian paternalism, utilitarism and justice. In C. Coons, & M. Weber (Ed.), *Paternalism. Theory and Practice* (pp. 216-230). Cambridge University Press. https://doi.org/10.1017/cbo9781139179003.012

Laibson, D. (1997). Golden Eggs And Hyperbolic Discounting. *The Quarterly Journal of Economics*, *112*(2), 443-478. https://doi.org/10.1162/003355397555253

Loewenstein, G., & Chater, N. (2017). Putting nudges in perspective. *Behavioural Public Policy*, *1*(1), 26-53. https://doi.org/10.1017/bpp.2016.7

Mills, C. (2013). Why Nudges Matter: A Reply to Goodwin. *Politics*, *33*(1), 28-36. https://doi.org/10.1111/j.1467-9256.2012.01450.x

Mills, C. (2015). The Heteronomy of Choice Architecture. *Review of Philosophy and Psychology*, *6*(3), 495-509. https://doi.org/10.1007/s13164-015-0242-7

Mitchell, G. (2004). Libertarian Paternalism Is an Oxymoron. *Northwestern University Law Review*, *99*(3,) 1,245-1,276. http://bear.warrington.ufl.edu/brenner/mar3503/articleslinks/libpat-oxy.pdf

Mongin, P., & Cozic, M. (2018). Rethinking nudge: not one but three concepts. *Behavioural Public Policy*, *2*(1), 107-124. https://doi.org/10.1017/bpp.2016.16

Nagatsu, M. (2015). Social nudges: their mechanisms and justification. *Review of Philosophy and Psychology*, *6*(3), 481-494. https://doi.org/10.1007/s13164-015-0245-4

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral And Brain Sciences*, *37*(1), 1-19. https://doi.org/10.1017/S0140525X12003214

Pettit, P. (1996). Institutional Design and Rational Choice. In R. E. Goodin (Ed.), *The Theory of Institutional Design* (pp. 54-89). Cambridge University Press. https://doi.org/10.1017/CBO9780511558320.003

Rebonato, R. (2013). A Critical Assessment Of Libertarian Paternalism. *Journal of Consumer Policy*, *37*(3), 357-396. http://dx.doi.org/10.1007/s10603-014-9265-1

Saghai, Y. (2013). Salvaging the concept of nudge. *Journal of Medical Ethics*, *39*(8), 487-493. https://doi.org/10.1136/medethics-2012-100727

Schmidt, A. T., & Engelen, B. (2020). The ethics of nudging: An overview. *Philosophy Compass*, *15*(4), Article e12658. https://doi.org/10.1111/phc3.12658

Schubert, C. (12 de octubre de 2015). *On the ethics of public nudging: Autonomy and Agency* (Working Paper 36). MAGKS Papers on Economics. http://doi.org/10.2139/ssrn.2672970

Simon, H. B. (1997). *Models Of Bounded Rationality: Empirically Grounded Economic Reason* (vol. 3). The MIT Press.

Smith, N. C., Goldstein, D. G., & Johnson, E. J. (2013). Choice without awareness: Ethical and policy implications of defaults. *Journal of Public Policy & Marketing*, *32*(2), 159-172. https://doi.org/10.1509/jppm.10.114

Sugden, R. (2017). Do people really want to be nudged towards healthy lifestyles? *International Review of Economics*, *64*(2), 113-123. https://doi.org/10.1007/s12232-016-0264-1

Sunstein, C. R. (2015a). Nudges, agency, and abstraction: A reply to critics. *Review of Philosophy and Psychology*, *6*(3), 511-529. https://doi.org/10.1007/s13164-015-0266-z

Sunstein, C. R. (2015b). Nudges Do Not Undermine Human Agency. *Journal of Consumer Policy*, *38*(2), 207-210. https://doi.org/10.1007/s10603-015-9289-1

Sunstein, C. R. (2017). People Prefer Educative Nudges (Kind Of). In *Human Agency and Behavioral Economics. Nudging Fast and Slow* (Palgrave Advances in Behavioral Economics book series. pp. 41-72). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-55807-3_3

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian Paternalism. *American Economic Review*, *93*(2), 175-179. https://doi.org/10.1257/000282803321947001

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.

Tversky, A., & Kahneman, D. (1974). Judgment Under Uncertainty: Heuristics And Biases. *Science*, *185*(4.157), 1.124-1.131. https://doi.org/10.1126/science.185.4157.1124

Uhlmann, E. L., Pizarro, D. A., & Bloom, P. (2008). Varieties of social cognition. *Journal for the Theory of Social Behaviour*, *38*(3), 293-322. https://doi.org/10.1111/j.1468-5914.2008.00372.x

White, M. D. (2013). *The Manipulation of Choice. Ethics And Libertarian Paternalism*. Palgrave Macmillan. https://doi.org/10.1057/9781137313577

Wilkinson, T. M. (2013). Nudging and Manipulation. *Political Studies*, *61*(2), 341-355. https://doi.org/10.1111/j.1467-9248.2012.00974.x