Abstract

Over the last two decades, advancements in cognitive and behavioral science have stirred lively debates in various academic disciplines on whether it is permissible for governments to use behavioral influences (so-called ‘nudges’) on citizens to improve their welfare. My dissertation shows why a careful moral consideration of behavioral influences goes beyond the standard nudge debate. I take up a broader approach which assesses whether cultivating and regulating behavioral influences for various purposes can be accommodated within the framework and principles of political liberalism. I call this approach behavioral enhancement. In the dissertation, I engage with three normative concerns at the core of behavioral enhancement: 1.) Under what institutional circumstances is it allowed for governments to nudge? 2.) Should the utilization of behavioral influences in markets be regulated? 3.) Should behavioral influences be used by government to get people to abide by enforceable moral duties?

The permissibility of government nudging, as well as the utilization of influences by market agents, is tested against the backdrop of liberal principles, of which personal autonomy is the most considered and explored here. I claim that this moral inquiry requires an account of personal autonomy updated by relevant considerations from the cognitive and behavioral sciences. I develop such an account under two empirical stipulations – pervasiveness of non-reflective behavioral influences and limited reflective resources of individuals. The account suggests that although intentional (as well as unintentional) behavioral influences have the capacity to undermine autonomy, they are also compatible with many individual management styles.

The first normative concern of behavioral enhancement engages us most with the standard moral debate on nudging. I particularly address the worry that nudges, as non-

transparent influences, cannot be reconciled with democratic principles of publicity and contestation, and cannot respect autonomy if they steer individuals without their consent. I develop a principle of ‘watchfulness’, which establishes institutional conditions for nudge transparency that allow individuals to accommodate nudges into their autonomous pursuits if they agree with them, or circumvent them without much burden if they do not.

The second normative concern starts with the observation that standard moral objections to government nudges should make us a lot more wary of influences by marketers. This is because these influences are not curtailed by principles of government nudging (mildness and sensitivity to agent preferences), and are more likely to overwhelm agents by virtue of sheer numbers. I lend further normative support to this observation, and recommend policy solutions for influences by marketers.

The third normative concern takes the first step in exploring the extent to which influences should be used to facilitate moral behavior. Here, I limit my advocacy of moral influencing on getting people to discharge duties that are either enforceable or non-enforceable due to feasibility constraints. I address worries that such influences promote mere conformity with duties, and that they stifle moral disagreement. The last chapter explores nudging to promote charity giving, which I claim is a case where the balancing of different duties – respect for autonomy and alleviation of poverty – is uncertain.

Introduction

In December of 2017, American economist Richard Thaler was awarded a Nobel Memorial Prize for his contribution to the field of behavioral economics. While Thaler started developing its main concepts in the early 1980s and published his first book in the field in 1992,[1] his most popularly resounding work is undoubtedly *Nudge: Improving Decisions About Health, Wealth and Happiness* (2008), co-authored with legal scholar Cass Sunstein. *Nudge* builds on a growing body of research from cognitive science and behavioral psychology, and proposes a number of tweaks in our choice environment that aim at rectifying systematic decision-making errors in humans, mostly those having to do with health and wealth. These tweaks the authors refer to as 'nudges'.

As an economist and a lawyer, Thaler and Sunstein draw their examples in *Nudge* primarily from their respective fields. Yet, it is apparent from the heuristics literature, which lies at the book's foundation, that cognitive error is a pervasive component of the human condition, and that all areas of human life are affected. This "diagnostic" side of the field, as opposed to Thaler and Sunstein's policy side, was developed by psychologists, led by Daniel Kahneman and Amos Tversky, who successfully pointed to a myriad of cases in which individuals systematically err due to their innate (and sometimes culturally conditioned) cognitive mechanisms. Kahneman's extensive work on cognitive heuristics earned him the aforementioned Nobel Prize in 2002. Several Nobel Prizes, for both the "diagnostic" and the policy sides of the new discipline,[2] show that behavioral economics is not merely a budding field, but has emerged a powerful social force that will spark academic and political interest for years to come.

[1] See Thaler, 'Toward a Positive Theory of Consumer Choice' (1980) and *The Winner's Curse* (1991).

[2] Nobel prizes were claimed by three other behavioral economists – George Akerlof (2001), Thomas Schelling (2005), and Peter Diamond (2010).

With Thaler's award in its collection, behavioral economics and nudging more specifically made their return into the limelight of popular attention, and the interdisciplinary craze for nudging is now likely to be rekindled. Yet, with its re-emergence on the academic stage, moral objections against the practice of nudging by governmental agencies have resurfaced as well. Most concern the manipulative aspects of nudging, the violation of personal autonomy, the lack of supervision over nudge designers, as well as the problems with paternalism and perfectionism. However, one objection that has received sparse attention in the moral literature asks whether using behavioral techniques can ever be part of democratic governance in liberal societies. It was raised recently by political scientist Henry Farrell (2017), shortly following Thaler's award reception. Such criticism is not the least bit surprising given the democratic crisis around the world, "post-truth politics", and the use of scientifically elaborated behavioral techniques in an attempt to sway public opinion.[3] In the wake of controversies and fears that covert means could further undermine the legitimacy of democratic procedures, we may legitimately ask whether the techniques that are administered by behavioral experts fall short of democratic standards.

Specifically, Farrell objects that because people are often nudged without being aware of it, a "nudgeocracy" lacks the pushback mechanisms that we can find in traditional democracies. If people are angry about being targeted by a nudge technique – either about its wrongly assessed aim, an aspect of the nudge itself, or simply because they resent being influenced by the government – it is harder for them to mobilize than in the case of standard legal regulation. Bad legal regulation, claims Farrell, provokes contestation and protest, and sometimes civil disobedience. Bad nudges go unnoticed, or are merely sidestepped (ibid.).

[3] Examples are certainly Hungary's ruling party Fidesz's use of guided questions in the 'National Consultation' survey seeking public opinion on immigration in 2017 (see, for example, Bearak [2017]) and recent controversies surrounding Facebook and Cambridge Analytica, alleging that personal identity information was purchased by politicians to influence voter behavior (see, for example, Rosenberg et al. [2018]).

Farrell’s concern regarding values of contestability and accountability highlights a dilemma surrounding the notion that has received some attention in the nudge literature – that of nudge transparency (Bovens 2009; Schmidt 2017). In short, the dilemma goes like this:

P1: We should accept only transparent policies (i.e., policies that can raise contestation and make governing bodies accountable).

P2: We should accept only effective policies.

P3: Nudges are effective only if they are non-transparent (or “in the dark” [Bovens 2009]).

P4: Nudges can be contested and accounted for only if they are transparent.

P5: All policies are either transparent or non-transparent.

C: Nudges are either effective, or they are transparent.

If correct, the argument points to a fundamental contradiction between the character of nudge policies and democratic principles. This is true, of course, if we assume that transparency is a necessary condition for democratic accountability and contestability, and if there is no use implementing nudges without their effectiveness on behavioral change. The argument shows that to nudge transparently is to nudge uselessly, and to nudge non-transparently is to undermine democratic principles. Such a sentiment among at least certain policy experts was confirmed by the Science and Technology Select Committee of the House of Lords in the UK, which concluded that, with regard to transparency and the basic conditions for democratic justification, nudges “involve altering behaviour through mechanisms of which people are not obviously aware. This raises an interesting question about the extent to which nudging is compatible with the Government’s commitment to ‘extend transparency to every area of public life’” (House of Lords 2011).

Nudge advocates object that 'nudging' is much broader than techniques that are 'effective only if they are non-transparent', i.e., techniques that owe their effectiveness to bypassing conscious reasoning. And since not all nudges have this effect, we can limit our endorsement to transparent nudge techniques. As Sunstein suggests: "If we value democratic self-government, we will be inclined to support nudges and choice architecture that can claim a democratic pedigree and that promote democratic goals" (Sunstein 2016a, 23). But as I explain later on, while these strategies may salvage parts of nudging as a policy project, the more narrow understanding of nudges as heuristic triggers is what raises the central moral objections, including that of Farrell, and it is these narrowly understood nudges that should be of main interest to moral and political philosophers, given their appearance as public manipulation.

My aim is to explore the moral properties of these controversial nudges and their place in a liberal democracy. Is it possible to use controversial nudges to guide and coordinate behavior in social settings, without undermining some of the basic principles of liberal democracy – openness, transparency, accountability and contestability?[4] Can nudges, in the narrow sense, be accommodated into a system of liberal democratic governance? And how does a system of safeguards affect the permissibility of using such behavioral influences, keeping in view the protection of citizens' personal autonomy? I will claim that, with certain limitations to the nudge project as defended by its main proponents, nudging can and should be accommodated into policy in a liberal and democratic institutional setting. Not only that, but the checks and balances that pertain to these democratic principles, I show, bring us closer to

[4] I reject, however, that nudges need to be compatible with democratic principles in all spheres of social life. For instance, I discuss in Chapter 5 that in cases where moral duties of citizens are hardly at all disputed, and in many cases codified legally, but in which citizens are failing to meet them, nudges can be used non-transparently. It would be sufficient, in such cases, in terms of democratic safeguards, that nudges can be contested by other behavioral experts.
I will be discussing the permissibility of government nudging, with the exception of Chapter 4, where I morally assess the behavioral influences utilized by marketers, and discuss whether a liberal democracy should find ways to curtail them.

resolving standard moral objections against nudging on grounds of accountability, epistemic defects, and autonomy.

Although Chapter 3 discusses in some detail how transparent nudging, if done right, can be part of democratic self-government, I do not draw on any specific democratic theory in this dissertation to ground the democratic principles that are meant to test the permissibility of nudging. I take it for granted that such principles, in the abstract, are at the core of most institutional regimes committed to liberal values. Publicity, for instance, a seasoned moral concept usually understood to tie political action not only to transparency but to socially shared reasons, can be defended on a number of moral grounds.[5] Thaler and Sunstein themselves appeal to Rawlsian publicity when they claim that a government should be banned "from selecting a policy that it would not be able to defend publicly to its own citizens" (Thaler and Sunstein 2008, 244).

They defend the principle on two grounds: first, publicity avoids the possibly volatile consequences of undisclosed information surfacing; second, without the readiness to defend policies in public, nudging becomes tantamount to lying, and thus treats people disrespectfully, as a mere means (ibid., 245). However, the concept still remains indeterminate. Is nudge publicity actual or hypothetical? Do nudge techniques have to be disclosed, possibly at the cost of their reduced effectiveness, or merely backed up by ends that could be justified by public reasons, in those rare cases where nudge designers are called out for them? If nudge publicity is merely hypothetical, then it is questionable whether it accomplishes the two aforementioned goals: first, a hypothetical publicity may easily fail to alleviate the volatility of disclosing secret information; second, it is not morally obvious that a hypothetical publicity treats people with

[5] For various understandings of publicity and their justifications, see, for example, Mill (1861) for his case against the secret ballot, Elster (1995) for his argument for "civilizing" political representatives, and Rawls (1996; 1999a) for his doctrine of public reason and public rules.

respect.[6] On the other hand, if nudge publicity is actual, then we re-encounter the problem of reconciling transparency and effectiveness, at least on the narrow understanding of nudges as heuristic triggers.[7] Finally, are the grounds 'defending nudge policies publicly' of a Rawlsian character, or are they derived from some other competing account of public reason?[8]

My account of nudge transparency should be sufficient, or so I will claim, for accomplishing Thaler and Sunstein's goals, as well as other liberal and behaviorally specific aims. My account enables citizens to actively participate in public affairs regarding the administration of behavioral techniques, and to contest these applications; it avoids a strictly top-down approach in which citizens do none of the nudge planning, and reconceives nudges as a means of socially coordinating self-regarding action. I aim to demonstrate how a nudge transparency is key for Thaler and Sunstein's initial goal – making people better off "as judged by themselves" (ibid., 5) – while giving dissenters all the tools for easily circumventing influences.

A positive case for nudging, however, curtailed by principles of easy resistibility and subjective benefit, only marks the beginning of a more serious appreciation of behavioral heuristics and influences in designing liberal democratic institutions. If nudging can be such a powerful force in the lives of citizens, then what of the powers of market practices to steer our lives? What of the capacity of influences to drive us into protecting others from harm, helping those in dire need, or promoting any number of important liberal values? What of the random psychological foibles that slow us down in the attainment of our goals? The efforts of

[6] This may depend on whether we only require a moral basis for mutual respect that reasonable individuals can endorse, or we are also interested in the citizens' *feelings* of disrespect when they find out they have been nudged. Thaler and Sunstein mention, for instance, that if citizens are told why they are enrolled automatically in a retirement program, they will not feel disrespected. But Thaler and Sunstein, like myself, operate in non-ideal circumstances, in which minimally informed citizens, upon realizing that someone played on their psychological foibles to achieve some desired outcome, may certainly feel disrespected.

[7] In later work (2013, 144-151; 2016a, 61-62), Sunstein does seem to suggest transparency and offering public reasons should be actual, rather than merely hypothetical.

[8] For one such account, see, for example, Laborde's concept of *accessible reasons* (2017).

cultivating behavioral techniques for a range of self-regarding and other-regarding purposes and institutionally regulating stifling behavioral influences together amount to a liberal project that I will call *behavioral enhancement*. This dissertation takes the first step in elaborating the normative questions of such a project.[9]

In the remainder of the Introduction, I quickly survey the nudge debate and introduce some of the issues that have been central and some that have received insufficient attention, while explaining how they relate to the normative questions of this dissertation. First, I provide more detail about the character of nudge techniques and the science that underpins them, and recount their recent prominence in policy. I then set up the theoretical foundations of behavioral enhancement in view of the most relevant moral concepts: political liberalism, libertarian paternalism, non-ideal theory, and anti-perfectionism. I finish with a chapter summary.

1.1. What are 'nudges'?

Nudges are notoriously hard to define. A possible starting point is offered by Jennifer Blumenthal-Barby (2013, 178), who shows that moral considerations of 'choice architecture' hinge on two phenomena. The 'bad choice phenomenon' refers to individuals predictably straying from decisions that achieve their ends, i.e., displaying a bounded rationality that biases them in their decision-making. The 'influence phenomenon', on the other hand, refers to the capacity of the choice environment to shape and steer choices in ways that circumvent conscious reasoning. To nudge, according to the standard understanding, is to cultivate this

[9] For constraints of space, this elaboration unfortunately cannot be completed here. Possibly the most relevant question that will be left unanswered is what kinds of influences should be permitted for us to have democratic equality and fair democratic procedures. Whether an ideal of 'behavioral neutrality' among contending agents and ideas in democracy is attainable carries great importance for illuminating a behaviorally enhanced liberal democracy and will, hopefully, be addressed in future work. Early considerations of behavioral influences in democratic theory can be found in Kelly (2012) and Ivanković (2016).

capacity in order to promote the well-being of those who are being influenced. In Thaler and Sunstein's words, nudging is "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" (2008, 6). In order to avoid forbidding options and changing incentives, Thaler and Sunstein's nudges must be "easy and cheap to avoid" (ibid.). Their claim, then, is that utilizing the influence phenomenon is permissible only if the influence is mild.

1.1.1. *Cognitive heuristics and biases*

That humans are boundedly rational is quite convincingly suggested by a mainstream theory in cognitive psychology – dual-process theory. Although there is some variation in terms of how the main tenets of the theory are laid out by different authors, which I explain in 2.2., all proponents agree on its basic axioms: there are two kinds of processes in the human mind, one fast, reflexive, associative, and operating at low-capacity, and the other slow, reflective, syllogistic and requiring exertion. Thaler and Sunstein (2008) and Kahneman (2011) use 'System 1' and 'System 2' for the two kinds of processes and show that although cognitive biases affect both, they are more commonplace in the processes of System 1. The two processes seem to run in parallel, but in many instances, one process takes over. Experienced drivers and professional athletes have developed their skills to the point that their actions in the relevant contexts require very little conscious effort (Sunstein 2013, 52). To disturb the balance of the two processes when much of the activity is learned and processed at low capacity is often detrimental to the performance of the activity. A player of a musical instrument knows all too well that thinking meticulously about every detail of his playing usually worsens it. The reflective processes, on the other hand, take over when we are solving a math problem, writing a doctoral dissertation, or performing unlearned tasks. Interestingly, speaking in a second language will be less prone to cognitive error, given the automatic and reflexive manner of speaking a native language, and a laborious and careful manner of speaking a foreign one (Keysar et al. 2012).

An important insight that establishes the relevance of this discussion is the realization of just how many cognitive heuristics seem to be contributing to the 'bad choice phenomenon'.[10] I only mention several here.

The status quo bias is the tendency of humans to pick options that they view as maintaining things as they are, i.e., a subtle preference for current states of affairs.[11] In debates on nudging, the status quo bias is commonly associated with the use of defaults, which determine what happens if people do nothing. In many countries, people's paychecks are automatically deducted for the amount of taxes they owe, while other countries endorse a 'do-it-yourself' tax scheme – between these, the first utilizes nudges to ensure tax abidance. From the heuristics literature, it is evident that the workings of the status quo bias go beyond mere inertia. One case study shows that when people weigh between several options to invest an inherited sum of money, they tend to pick options that are framed as maintaining the way in which the money was invested before they inherited it (Samuelson and Zeckhauser 1988).

Humans also postpone acting on commitments and obligations, often dangerously close to approaching deadlines. Procrastination is a self-control problem that concerns a gradual change in salience of costs and benefits for a certain activity over time (Akerlof 1991), as the costs begin to loom larger the closer we are to a deadline. This often produces painful consequences of failing to meet the deadline, or performing the task below an expected quality.

[10] In Chapter 2, where I explore how the fact of the boundedly rational mind is squared with philosophical conceptions of personal autonomy, I point out that System 1 heuristics should not be viewed merely as an evolutionary blunder in our cognitive wiring. Most of the time, heuristics are effective cognitive shortcuts that allow us to economize reflective resources. The 'bad choice phenomenon' only refers to cases in which these evolutionary mechanisms fail. I later discuss how this contributes to a more psychologically developed account of autonomy.

[11] The bias intersects with other heuristics, such as loss aversion and the endowment effect (Kahneman et al. 1991).

Oftentimes, people are successful in alleviating the effects of procrastination, by imposing themselves with deadlines, but these are not as effective in improving task performance as externally imposed deadlines (Ariely and Wertenbroch 2002).

Anchoring occurs when we base our estimations regarding a certain value on some other value that is mentioned at the start of the estimation (Tversky and Kahneman 1974, 1128-1130). Consider the success of using discounts on goods in stores. The former prices represent strong stimuli for people to buy, regardless of whether they truly relate to discounted prices. But anchoring need not only bear on estimations of amounts. In a study that pays homage to a question-ordering experiment from the 1950s, test subjects from the US were asked two questions. The first asks whether the US should allow Eastern Bloc reporters free entry and reporting. The second asks whether US reporters should be allowed to freely report in Eastern Bloc countries. The results show respondents were more likely to allow both US and Communist reporters to freely report if the question about American reporters was posed first. If that question is answered in the affirmative, their response to the other question was driven by reciprocity. If the question about Eastern Bloc reporters was posed first, and the respondents say ‘no’, their response to the other question was more likely to be negative as well, in this case being driven by consistency. In both cases, the answer to the question posed second was anchored in the response to the question posed first (Schuman and Presser 1981).

Humans are also susceptible to how choices are framed. This insight brings into question the assumption of invariance in rational choice theory, which states that human preferences are stable across different presentations of choices with equal option sets. One of the most oft-mentioned examples in the nudge literature is from a study by McNeil et al. (1982), which shows that patients respond differently to choices regarding therapy with the same probability values, depending on whether the framing emphasizes the probability of living, or of dying. Another study (Ubel et al. 2010) shows that on one version of framing effects – the order effect

– patients are more likely to take a medication if its risks are presented first and the benefits second.[12]

Loss aversion, and its heuristic correlate, the endowment effect, are yet other strong drivers of less-than-rational behavior, which show that individuals are more averse to losses than they are drawn to gains that are equal in value. People suffer more from losing, say, a hundred dollars, than they extract pleasure from gaining the same amount; in fact, the gains may have to exceed the amount significantly to outweigh the disutility of the loss (Kahneman and Tversky 1992). These cognitive biases may manifest not only as material losses. They explain why we resist admitting that certain activities were a waste of time and effort, and why we often stick with them even after losses become salient.

This short list of cognitive heuristics and biases is by no means exhaustive.[13] It only describes the heuristics that have been most extensively explored in behavioral studies, and helps to paint the picture of the human mind that deviates from the rational ideal of classic economic theory, and, as we shall see in Chapter 2, poses interesting challenges to standard philosophical conceptions of personal autonomy.

1.1.2. ***The elusiveness of nudges***

The last decade has seen an explosion of policy attempts by institutions and their corporate partners to exercise influence over individuals in a range of areas. Possibly the most famous nudge, the cafeteria food arrangement, operates via salient visual cues to prompt healthier food choices. The urinal fly, first tested at Amsterdam's Schiphol Airport, is also a visual cue that significantly reduces spillage in men's toilets (Thaler and Sunstein 2008, 3-4). As mentioned earlier, enrolling people automatically into retirement programs predicts higher

[12] A wide variety of framing kinds can be cultivated as part of the influence phenomenon, including equivalency framing, emphasis framing, question-ordering effects, and question-wording effects. For detailed descriptions, see Kelly (2012, 12-18).

[13] For a more extensive list, see Kahneman et al. (1982), Kahneman and Tversky (2000), and Gilovich et al. (2002).

participation numbers, by virtue of the status quo bias. Framing effects are used for steering people towards choices that their physicians deem supportive of their aims. Social norms are used to promote pro-social behaviors and counter tax evasion.

Yet, the general character of nudge techniques remains elusive. I explicate several reasons here. The first is that, when nudge techniques are scanned for common features, the 'nudge' category turns out to be remarkably heterogenous. The heterogeneity is due to a variety of ways in which cognitive heuristics steer decision-making, and the different designs that to varying degrees allow persons to consciously interact with influences. Consider the following two influences. The so-called 'decoy effect' is commonly used by restaurants and other companies that arrange price lists for their products. For instance, a restaurant could offer a less expensive and a more expensive lunch deal. Adding a decoy in the form of a third, most expensive lunch deal, makes it more likely that the more expensive of the previous two will be sold. Now compare this to organ donation regimes, which in many countries use opt-out defaults to boost participation numbers. Both techniques are tapping into less-than-rational cognitive processes, but it is not quite obvious that the influences are similar enough in kind. The exploitation of the status quo bias does not obviously raise the same normative qualms as the decoy effect or the exploitation of some other heuristic.

While this might be coined the 'qualitative problem', the second reason points to the quantitative dimension of nudge techniques. Recall Thaler and Sunstein's qualifier that nudges are to be "easy and cheap to avoid" (2008, 6), implying that the nudge project only vouches for mild influences. But what exactly is the measure of mildness? How do we assess whether conditions for avoidance have been established? In Chapter 3, I suggest that certain influences are indeed more difficult to avoid than others, even when made transparent. Still, these piecemeal insights are not likely to add up to a scale of influence strength. Governments might end up doing more than nudging, given the lack of a reliable measure for ruling out stronger influences.

Why are behavioral sciences having trouble solving the qualitative and quantitative problems? Till Grüne-Yanoff offers a compelling explanation. He posits that behavioral studies are able to show "*how much* the policy intervention, in a particular environment, makes a difference" to behavior, but not "*how* – through what processes or mechanisms – the intervention produces this behavior" (Grüne-Yanoff 2016, 465; emphases in the original). The argument shows that behavioral sciences are only able to observe "inputs" and "outputs", i.e., changes in the choice environment and the resulting differences in behavior, as compared to the behavior of control groups. These findings reveal very little about the underlying cognitive processes and only hint at how the division of cognitive labor is structured in the human mind.

Is Grüne-Yanoff's argument too strong? One objection might be that the solution is fairly simple: we should stipulate that influence strength is determined by the impact of the intervention on behavioral change in a population over time. But this solution would be too simplistic. As Chapter 3 shows, some interventions seem to have considerable impact on behavior, but only while they are undisclosed, whereas some interventions have a lasting effect even if disclosed. Yet, it is perfectly within realms of contingency that the techniques with greater impact on behavior are the more avoidable ones when disclosed. This is not to suggest that behavioral influences will never be quantifiable in some sense, but it seems influence strength will have to be tested further in studies that enable subjects to spot the influences and consciously process them. Such evidence is currently scant.

The third reason goes back to Blumenthal-Barby's influence phenomenon – the capacity to steer choices in non-conscious ways by changing the choice environment. In many examples in the nudge literature, however, the 'non-conscious' aspects of influences are nowhere to be found. According to this *broad* conception of nudging, any alteration in choice architecture that predictably results in behavior change and leaves options open counts as a nudge, regardless of whether it bypasses reasoning. This conception is endorsed in *Nudge* and in all of Sunstein's subsequent books on nudging (e.g. Sunstein 2013; 2015b; 2016b). Just how much this move broadens the concept is evidenced by the inclusion of interventions such as providing "clear, simple information about healthy diets", a "disclosure requirement imposed on providers of retirement plans, so that people can clearly see their projected monthly income in retirement", or a "reminder, by telephone or text, that a consumer is about to go over his or her allotment of monthly minutes" (Sunstein 2013, 42). Similarly, Sunstein often mentions the GPS as an example of nudging.[14] The broad conception makes it difficult for us to differentiate nudging from simple information giving or offering reasons for action. Interventions like warnings and reminders, as well as the GPS, a tool for navigation, are obviously benign techniques that hardly compare to morally controversial instances of nudging.[15] In my opinion, by pointing to the morally innocuous examples, nudge advocates are able to water down the criticisms to the less innocuous techniques of the nudge project, as well as emphasize just how unavoidable nudging is, making the "antinudge position […] a literal nonstarter" (Thaler and Sunstein 2008, 11). But the conflation of nudging and mere information giving does more harm than good in conceptual terms, and in the very least calls for differential moral treatment of 'nudging as information giving' and 'nudging as heuristics triggering'.

Granted, the distinction is not always a foolproof diagnostic tool in practical cases. One reason is certainly the ubiquity of framing effects, which entails that decisions may vary depending on the presentation of information; since our option-weighing might well depend on how often a certain piece of information is brought up or how it is framed, there may be less-

[14] See, for example, Sunstein (2015a, 512; 2016a, 26, 36).

[15] As I shortly go over in 1.2.2.2., the moral debate on nudging has most often been interlocked with the one on paternalism. In the latter debate, most authors endorse by default that persuasion and information giving cannot be paternalistic. A notable exception is the view offered by Tsai (2014).

than-rational aspects to information giving. Still, these worries do not leave us conceptually incapacitated. In most cases, we have a good idea which nudges bypass reasoning and which aspects of the nudges account for it.[16]

Throughout this dissertation, I will explore the normative questions surrounding a narrow conception of nudging and other behavioral influences, namely, techniques that result in motivational modulation and "heuristics-triggering" (Barton and Grüne-Yanoff 2015, 343). Yashar Saghai offers a useful definition of the narrow understanding of nudging: "A nudges B when A makes it more likely that B will φ, by triggering B's automatic cognitive processes, while preserving B's freedom of choice" (2013, 487).[17] Differentiating between certain cases of heuristics triggering and mere information giving remains problematic, but the solution hardly lies in the concept's uncontrolled expansion. Warnings, reminders, GPSes and other kinds of information giving are normatively mundane and uncontroversial, and have been used (primarily by Sunstein) as a rhetorical device to deflect from the morally dubious side of the nudge project. Should we organize parts of our individual and collective lives by appointing experts to steer our behavior via non-conscious stimuli? When, if ever, is the utilization of such techniques permissible? And if it is not, what about the plethora of such influences by non-governmental agents which already steer our supposedly autonomous lives? These are the questions that should be mulled over by ethicists and political philosophers.

Note that I also use a narrow conception of other behavioral influences that are not nudges in the standard sense – namely, those not curtailed by resistibility and the principle of

[16] There are other reasons for nudge elusiveness that deserve honorable mentions: 1.) In order to add to the relevance of nudging as a new mode of policy making, many methods are included under its banner that do not neatly fit the mold. Most obviously from its definition, nudges are not supposed to change economic incentives, but many techniques do just that. One such example is the incentive for people in Malawi to pick up their HIV test results for one tenth of their daily income (Institute for Government and the Cabinet Office 2010, 20); 2.) 'Choice architecture' may not be attached to any particular instance of choice. For instance, a particular information frame or the ordering of news can simply have a lasting effect on how we contemplate a certain subject; 3.) As Chapters 5 and 6 will show, nudges could be given justifications that are not 'self-regarding'. The benefits of many nudges, like the urinal fly or opt-out organ defaults accrue not to nudgees, but to other people.

[17] See also Heilmann (2014, 79).

benefiting individuals by their own lights. This particularly concerns the majority of influences I discuss in the second part of the dissertation. Yet, for reasons of brevity and simplicity, I will also refer to these influences as ‘nudges’. ‘Nudge’ will be used here as an umbrella term for any heuristic-triggering behavioral influence, that is, any influence as conceived in the narrow sense. Chapters 4 and 5 will elaborate the usage of the term further in their separate normative considerations.

1.1.3. *The popularity of nudge units*

My focus on nudges in the narrow sense means that my critique will only capture a subset of techniques researched by behavioral units. My primary focus is on whether heuristic triggers are compatible with and can be cultivated by liberal democratic government, not on how behavioral units have been conducting their business so far. Still, I should say a few words about just how prominent nudging has become among current public institutions. The trend should indicate the urgency of our ethical treatments.

Shortly after its emergence, nudging has been granted the status of "one of the hottest ideas in current policy debates" (Hausman and Welch 2010, 123), and has since inspired normatively ambivalent visions of a "nudge-world" (Waldron 2014) and a "Republic of Nudges" (Rachlinski 2017). It is slowly, but surely, gaining traction in the policy world, especially for the success of nudge-specialized units in cutting down administrative expenses during the economic crisis of the late 2000s. So far, offices specialized for testing and applying behavioral insights have been opened in the US, the UK, Australia, Canada, the Netherlands, Germany, South Korea, and Denmark, as well as in international institutions like the World Bank and OECD.[18]

The beginnings of nudging's policy successes, following *Nudge*, probably came with Sunstein's appointment to the Office of Information and Regulatory Affairs (OIRA) at the White House during Obama's presidency. During his time as OIRA's administrator (2009-2012), Sunstein oversaw "nearly two thousand rules from federal agencies" across a vast range of policy areas (Sunstein 2013, 9-10). OIRA adopted nudging in a successful attempt to curb budgetary costs and sensitize policy makers to scientific rather than anecdotal evidence.

[18] Somewhat surprisingly and in spite of criticisms about the growing size of its bureaucracy, institutions of the EU have not yet taken a big interest in behavioral techniques. One notable exception is the recent EU cookie directive, which requires explicit consent from Internet users. See European Parliament and European Council (2018).

However, Sunstein's contributions to OIRA met criticism from both the left and the right, in spite of his conviction that nudging appeals to both sides of the political spectrum. In Sunstein's own words, the left seemed more interested in strict mandates (ibid., 26).[19] Bernie Sanders, for instance, accused Sunstein of helping big banks by not regulating them (ibid.: 30-31). On the right, conservative Glenn Beck condemned nudging particularly for its secretive character, and labeled Sunstein "the most dangerous man in America" (ibid., 28-29).

The first office specialized for the application of behavioral sciences was the UK's Behavioural Insights Team (BIT), founded in 2010. According to Adam Burgess, BIT's contributions in the UK were thought to be "in the context of a wider devolution of power to local communities" (2012, 5). BIT established ties with both governmental and private sector agencies, and produced a great number of publications recommending policy solutions in the areas of charity, decreasing fines, tax abidance, electoral participation, energy consumption and sustainability. It has become the most animated nudge-specialized body, working with local administrations and setting up offices in the US, Australia, and Singapore.

The approval rate of nudge techniques is also fairly high across very diverse world countries. A study by Hagman et al. (2015) in Sweden and the US shows that even the techniques with the lowest support are approved by more than 50% of respondents. More strikingly, the study shows that the approval rates are considerably higher for other-regarding nudges (benefits accrue to other people, like organ donation or environmental nudges) than self-regarding nudges (benefits accrue to the individual, like health-inducing nudges). Another study, by Sunstein et al. (2018), shows that citizens of Western and Far Eastern countries widely approve of nudges, but are more likely to reject techniques that are found manipulative. These

[19] Nudging is often discussed as an anti-regulatory method. This is understandable, given its rise in circumstances of austerity. However, there is nothing conceptual about nudging that would make its advocates anti-regulatory and right-wing. In the literature, nudges are rarely expected to overtake regulation wholesale. As Chapter 5 shows, nudges can often be used to facilitate complicity with codified regulation if the latter is ineffective. It should not be assumed that democratized nudging is meant to oust coercive regulation.

findings are interesting for two reasons. First, they show that nudge units are here to stay, given that they are widely endorsed by governments and citizenry alike. Second, they show that citizens still have reservations for the kinds of nudges that are here analyzed as normatively controversial.[20]

1.2. Theoretical foundations

1.2.1. ***Political liberalism***

I have noted that the regulation and cultivation of behavioral influences would be pursued and defended in this dissertation under the purview of liberal democracy. The theoretical legacy is inherited here from the 'high liberalism' of John Rawls. Political liberals of a Rawlsian ilk have focused on refining the basic principles of a just society, which would deliver the conditions for citizens to cooperate and coordinate their efforts.

I will specifically consider how behavioral facts bear on Rawls's two moral powers, which people are presumed to have as free and equal citizens in a fair system of social cooperation. First, citizens are presumed to have "the capacity to understand, to apply, and to act from the public conception of justice" (Rawls 1996, 19). Second, they possess "the capacity to form, to revise, and rationally to pursue a conception of one's rational advantage or good" (ibid.). Now let us also assume, in a standard liberal vein, that citizens are indeed able to form considered judgments about the public conception of justice and their conception of the good. Behavioral facts should still, at least in many circumstances, produce obstacles for individuals to act from the public conception of justice, or to revise and pursue their conceptions of the

[20] However, the category of 'manipulative nudges' that Sunstein et al. examine might be narrower than heuristics-triggering nudges. Respondents primarily rejected subliminal messaging and visual illusions to prevent speeding. It is also not certain from the study whether respondents appreciated the effects of behavioral heuristics. The hypothesis that overt influences are more acceptable to respondents than covert ones is more strongly suggested in Felsen et al. (2013).

good. Rawls himself seemed to have been acutely aware of these obstacles, and argued for a society-wide commitment to overcome them:

> “It is […] rational for [parties in the original position] to protect themselves against their own irrational inclinations […] by accepting certain impositions designed to undo the unfortunate consequences of their imprudent behavior. For these cases the parties adopt principles stipulating when others are authorized to act in their behalf and to override their present wishes if necessary; and this they do recognizing that sometimes their capacity to act rationally for their good may fail, or be lacking altogether” (Rawls 1999a, 219).

Rawls resumes:

> “Thus the principles of paternalism are those that the parties would acknowledge in the original position to protect themselves against the weakness and infirmities of their reason and will in society. Others are authorized and sometimes required to act on our behalf and to do what we would do for ourselves if we were rational, this authorization coming into effect only when we cannot look after our own good. Paternalistic decisions are to be guided by the individual’s own settled preferences and interests insofar as they are not irrational, or failing a knowledge of these” (ibid.).[21]

It is Rawls’s understanding of commitment to the avoidance of irrationality that drives my account of nudging, as well as the regulation and cultivation of influences in a wider sense that amounts to behavioral enhancement. As I will show throughout, nudging and regulating influences can be an expression of this commitment, provided that there are democratic safeguards in place which ensure that influences are contestable, and that nudgers have at their disposal the information to pursue just and rational ends. These influences and regulations will help citizens to sustain the moral powers which would otherwise be compromised, that is, to act from a public conception of justice and authentically pursue their conceptions of the good.

Note that the commitment of parties in Rawls’s original position to counteract their own cognitive failures does not seem to be limited only to whether these failures are exploited by others, like sinister nudgers or profit-seeking marketers. Rather, the principle allows

[21] A conceptual disagreement might arise between myself and Rawls about whether these principles are truly ‘paternalistic’ if they would be acknowledged and accepted by parties in the original position, given the standard understanding of paternalism to be against the will of its targets. It is possible that Rawls retains the concept of paternalism assuming that hypothetical consent in the original position is "distant" from the consent we give in actualized social settings. Still, I believe that sticking to the term bears little substantive weight for Rawls. I will discuss the matter of paternalism more thoroughly in the following subsection.

government to intervene when individuals fail to act on their settled preferences, or exercise their moral powers, even when this is not brought about by the actions of others. Not all liberals would agree with this kind of normative mission for the state. Isaiah Berlin, for instance, believed that we are free insofar as we are "unobstructed by others", and that the task of the liberal state is to prevent such obstruction between citizens (Berlin 1969, 122). Thus, the Rawlsian commitment to protect citizens from their own irrationalities, that I endorse here, goes beyond the basic functions of the liberal state.

However, the commitment is, I believe, easily understandable within the context of the Rawlsian project. Not only is the exercise of the two moral powers at the core of understanding citizens as free and equal, but the differential advantage that citizens might attain in a liberal society that results from better natural assets (for instance, mental powers to overcome psychological frailties) is, according to Rawls, "arbitrary from a moral point of view" (1999a, 274), given that citizens do not deserve such assets. The inequalities that arise from some individuals exercising better self-control or being able to hire others to make more sound judgments for them would not be justified from the point of view of the difference principle.[22] Thus, if it were within the state's capacity to help others to overcome psychological frailties in order to exercise their two moral powers, then such capacity should be employed.

Despite different takes on the extent to which liberal states should intervene, most liberal views would agree that something like the two Rawlsian moral powers should be protected and enabled. Insofar as my normative project here is guided by the two moral powers, it earns its liberal pedigree. But while protecting and enabling the two moral powers, and thus, personal autonomy, is given particular emphasis in this dissertation, autonomy is one weighty consideration among others. Specifically, personal autonomy will not be treated as a *side-*

[22] Rawls's difference principle states that distributive inequalities are justified only if they benefit the worst-off members of society. See Rawls 1999a, specifically 65-70.

constraint, in the sense that it will not fully restrict what may be done to persons in the face of more valuable pursuits. Several times in this dissertation, I will conduct a consequentalist weighting of valuable options, among which autonomy is a particularly important consideration, but not always decisively so.[23]

Consider harm, another weighty and liberally acknowledged consideration. I give special attention to the cultivation of influences against harm in Chapter 5. Some liberal values may be pursued via nudges, or the regulation of behavioral influences, even within the Rawlsian project specifically. The budding field on implicit biases against stigmatized social groups,[24] for instance, and the behavioral methods of ousting them,[25] bear particular weight on Rawls's argument that public offices and occupations need to be open to all under conditions of fair equality of opportunity (1996; 1999a). Even the difference principle can be promoted via "a Rawlsian nudge", as exemplified by Jaime Kelly, who states that a referendum on property tax will more likely be supported if it is framed in terms of 'maintaining previous increases' than 'raising taxes' (2013, 223-224). In time, the liberal politics of behavioral enhancement will have to incorporate the pervasiveness of behavioral influences into more detailed considerations regarding these and other liberal values, but for now, I leave such considerations for future work.

1.2.2. ***Beyond the fixation on libertarian paternalism***

On Thaler and Sunstein's account, the practice of nudging is espoused to their position of libertarian paternalism (hereinafter, 'LP'). Nudging is paternalistic, they believe, because "it tries to influence choices in a way that will make choosers better off, *as judged by themselves*",

[23] I will also work with the assumption that autonomous pursuits can sometimes be worthless. Individuals may use their autonomy for pursuits that are obviously and gravely wrongful. In such circumstances, other values will gain the upper hand. But the worthlessness of such actions does not render them non-autonomous. These assertions will be particularly important in Chapters 5 and 6.

[24] For an overview of important considerations surrounding implicit biases across philosophical fields, see Brownstein (2015), and for a psychological elaboration on the malleability of implicit biases, see Dasgupta (2013).

[25] See, for instance, the elaboration of the 'evaluation nudge' in 5.1.3. (Bohnet et al. 2015).

but a paternalism with nudging at its core "is a relatively weak, soft, and nonintrusive type of paternalism because choices are not blocked, fenced off, or significantly burdened" (Thaler and Sunstein 2008, 5). I claim that the debate on nudging has so far been needlessly fixated on this concept. For moral philosophers in particular, who invest a lot of energy into conceptual clarity and precision, the term has been particularly distracting. The permissibility of nudging should instead be discussed in broader normative terms.

One strategy for rejecting LP is to claim that the terms themselves do not fit – that the concept is oxymoronic. Scrutinizing LP, Gregory Mitchell argues that libertarianism and paternalism are at odds, and that it is libertarianism that gets the shorter end of the stick (2005, 1247). My strategy here is to point to the arguments showing that the nudge project commits to both libertarianism and paternalism in a very thin sense. Divorcing nudging from LP, as well as libertarianism and paternalism separately, can help us to reset the debate and approach behavioral influences freely from certain ideological overtones.

1.2.2.1. *Libertarianism*

LP's appeal to libertarianism seems to rest primarily on preservation of liberty (Thaler and Sunstein 2008, 5). Here I look at three arguments suggesting that LP is not truly libertarian.

One line of criticism should be that non-transparent influences safeguard liberty only in a weak sense. Take the defaults automatically enrolling individuals into organ donation programs. Libertarian paternalists point out that individuals can opt out at low cost, but, in practice, they rarely do, and the reverse would likely be true in case of an opt-in default. Choice architects explicitly design choice contexts with such predictable outcomes in mind. In what way is LP then committed to preserving liberty? It would seem that it allows that decisions be heavily biased against certain options. And given that cognitive science is yet to solve the quantitative problem, the cognitive pull might leave alternative options only nominally free.

Libertarians have yet to explain how the fact of behavioral influences fares against the notions of negative rights and voluntariness, or how strong influences compare to milder cases of coercion (like small financial penalties). It is likely that in choosing between different choice designs, a truly libertarian choice architect would be cautious to avoid influences that could undermine negative rights and voluntariness. In the organ donation case, the architect would not be without alternatives. Consider prompting drivers to indicate their donor preferences on their driver's license. The libertarian could grant that such prompted choices are not devoid of behavioral influences (e.g., in the form of framing effects), but maintain that they more convincingly respect negative rights and safeguard voluntary choice than automatic enrollment with the possibility to opt out. The true libertarian could, thus, still convincingly compare at least certain designs in terms of which is more likely to undermine negative rights and voluntariness.

Secondly, libertarians would claim that a minimal state is only allowed to protect negative rights, not promote welfare. As Richard Arneson states, libertarianism "include[s] no rights to be given positive assistance, aid, or nurturance by others" (2000, 41). This libertarian commitment would not rule out at least certain nudge policies (such as prompting the expression of organ donation preferences), namely those that would preserve an individual's voluntary choice. Still, the libertarian would reject policies that aim at welfare gains for individuals and the society. Since Thaler and Sunstein's LP is a markedly welfarist position, it is not truly libertarian (Mitchell 2005, 1260-1264).

A final suggestion might be that LP is libertarian only in that it aims to downsize the government's coercive apparatus. However, first, it is not obvious that libertarian paternalists make such a commitment. True, nudgers have made their careers on saving money for governments, but this is not the sense in which libertarians are primarily interested in downsizing government. It is perfectly contingent for nudge policies to cut budgetary expenses while preserving the same level of coercive control. Second (as I explained in footnote 19), nudges are often considered substitutive to coercive regulation. Yet, there is nothing conceptual about nudging, or within the main tenets of LP, that would suggest nudging serves the libertarian purpose of achieving a smaller state. The nudge state and the libertarian state might end up matching each other in the size of their coercive apparatuses, but that does not suffice to call LP properly libertarian.

1.2.2.2. *Paternalism*

Gerald Dworkin defines paternalism as “the interference of a state or an individual with another person, against their will, and justified by a claim that the person interfered with will be better off or protected from harm” (2002). Here, I test whether the theoretical assumptions of libertarian paternalists are appropriately labeled paternalistic. Not all techniques will be considered. The broad conception of nudging lists warnings and information giving, measures that obviously do not count as paternalistic on almost any philosophical reading.[26] I focus only on techniques of the narrow conception. I will claim that with an easy opportunity to consent to or reject the guidance of techniques, which I will argue for in Chapter 3, it is doubtful that most nudges are truly paternalistic.

Looking at the range of available nudges, we notice that paternalistic justifications are not joined with many of their effects, or at least that such justifications are not primary. For many interventions, as I will show throughout the dissertation, it could be rightfully claimed that benefits are split between the targeted individuals and other members of society. For some of these at least, other-regarding considerations clearly outweigh self-regarding ones, as well as that many interferences are not backed in the least by paternalistic reasoning. The default switch in organ donation schemes, in most cases, is obviously directed not at helping would-be

[26] As I already mentioned, one exception is Tsai (2014).

donors, but those in need of functioning organs. Instead of employing an influence for self-regarding reasons, the government may try to alleviate some grave harm or make sure that citizens act in line with enforceable principles of social justice. For instance, in the case of the cafeteria food arrangement, typically discussed in the context of promoting individual health, individuals may also be nudged due to the nudger's intention to reduce health care costs, or lower their carbon footprint (Mills 2018, 397). Of course, self-regarding justifications are central in other cases, or relevant enough (as in the cafeteria food arrangement) not to be ignored. But the simple fact of many techniques for which other-regarding considerations bear more weight, being rooted in what citizens owe each other, points to the oddity of nudge debates being fixated on paternalism.[27]

Sunstein also envisions LP to be a *means* paternalism, rather than an *ends* paternalism. Means paternalists are only interested in influencing the means with which people achieve *their own* ends. Ends paternalists, on the other hand, aim to direct people towards ends which are chosen by the paternalists for them, for instance, health and wealth broadly considered (Sunstein 2015b, 61-63). A paternalism aiming strictly at influencing means often runs into feasibility problems. This is because individuals in society live in a wide pluralism of ends, and nudge policies, which in most cases have more than a single target, are not tailored for each set of ends separately. This is why nudges more often appeal to presumed and generalized ends of savers rather than spenders, the healthy rather than the unhealthy, and to the risk-averse rather than risk seekers. But we should stop to consider that if we were to overcome these feasibility constraints, there would be very little 'paternalism' left. Sunstein acknowledges this himself:

> "We should be able to agree that government would focus only on means, and indeed would not be paternalistic at all, if it could have some kind of access to every person's internal concerns

[27] Paternalistic reasons, Grill (2007) convincingly argues, need not be primary for the relevant interference to be qualified as paternalistic. It would be sufficient that some significant aspect of the conjunction between the reason and the action is paternalistic for the interference to have a paternalistic flavor. However, it is perfectly conceivable for some popular nudges, usually assumed to be paternalistic, to be grounded entirely in non-paternalistic reasoning, and with paternalism as a mere afterthought.

> and provide them with accurate information about everything that already concerns each of them. Perhaps in the fullness of time, government or the private sector will be able to do something like that. But insofar as government is being selective, it is at least modestly affecting people's ends, perhaps even intentionally." (ibid., 67-68)

Why is a fully attained means paternalism here not 'paternalism' proper? Presumably, the attainment of means paternalism would be an indication that society has found ways of communicating the ends of citizens to the nudging agencies of the government, and that their nudge techniques affect only the citizens with fitting ends. Assuming that governments could also find out how much of their interference would be acceptable to the individuals in question, a central feature of paternalism would effectively be eliminated – that it is 'against the will' of targeted individuals (Trout 2009). If this is the case, LP is not a paternalism *in principle*, but only due to feasibility constraints. An account of nudge transparency for self-regarding considerations, which I offer in Chapter 3, fits the purpose of overcoming these feasibility constraints. If nudgees can see nudges coming and circumvent them in accordance with their own ends, then they can consent to nudges, dissolving the worry of paternalism. Insofar, my account of nudging is only paternalistic if it is unfeasible. As Joel Anderson says, "if the nudges really had the consent of those being nudged, it would no longer be clear why the approach would need to be called 'paternalistic' at all" (2010, 374).

One final consideration is that LP might be paternalistic because of an attitude of superiority and disrespect assumed by governments and nudge experts. Nudgers are paternalistic, the argument goes, because the practice of nudging is an expression of cognitive superiority. However, such an attitude need not be present. The "superiority" of nudgers is owed to the design level on which they are not prone to the cognitive foibles that occur when they engage with day-to-day decisions. As Andrés Moles states, it is not based on comparative judgments regarding people's abilities to pursue their ends (2015, 652-653). Nudging, therefore, is not disrespectful.

1.2.3. ***Non-ideal theory***

Another foundational consideration inherited from the Rawlsian legacy is how idealized our theorizing about political institutions and agents is. According to Rawls, to engage in ideal theory is to stipulate strict compliance by agents and favorable circumstances in which a well-ordered society can be maintained (1999a, 216). Ideal theory takes "men as they are and laws as they might be" (1999b, 7). Non-ideal theory observes the obstacles to the two conditions of ideal theory – full compliance and favorable circumstances – and assesses the permissibility, as well as the feasibility, of institutional arrangements and policies *en route* to a well-ordered society (ibid., 89). I use Laura Valentini's three interpretations of the ideal/non-ideal distinction (2012) to explain three ways in which my approach in this dissertation is non-ideal: 1.) moral agents are partially compliant as opposed to fully compliant; 2.) circumstances are realistic rather than utopian; 3.) normative assessment is transitional rather than end-state.

First, the partial compliance of moral agents is stipulated in virtue of the pervasive non-reflective influences on human behavior. Given that cognitive heuristics are an evolutionary trait, I take partial compliance brought about by behavioral factors to be a fairly permanent state of human affairs. The principles with which individuals fail to comply include natural duties such as "not to harm or injure another" or to provide "mutual aid" (Rawls 1999a, 98), as well as "to support and to comply with just institutions that exist and apply to us" (ibid., 99). Much of my normative theorizing in this dissertation will concern, as John Simmons calls it, the unfortunate inability of agents to comply, as opposed to deliberate non-compliance (2010, 16-17). In such circumstances, my account will detail the means that would get agents to act in line with their own moral judgments. However, this does not entail that my account overlooks deliberate non-compliance. Particularly in Chapters 5 and 6, I ask whether the deliberately non-compliant agents can be exposed to behavioral influences that are meant to bring them in line. It might be suggested that accounting for the effects of cognitive heuristics and choice environments does not obviously plant us into non-ideal theory – Rawlsian ideal theory includes "[t]he general facts of moral psychology" (Rawls 1999a, 126), or, as previously mentioned, 'men as they are'. Still, Rawlsian theory says little about how, in the absence of 'laws as they might be', the psychological features discussed in this dissertation may contribute to compliance being partial instead of full. This suggest that non-ideal conditions regarding compliance might come hand in hand with non-ideal institutions.

This brings me to the second dimension in which my theorizing is non-ideal – realistic, as opposed to utopian circumstances. In Rawlsian theory, circumstances are favorable if they can sustain a well-ordered society, that is, a perfectly just liberal democracy that can "come about and be made stable under the circumstances of justice" (2001, 13). My account departs from this ideal, both in the domestic and international spheres; due to a lacking social technology, societies fail to produce domestic and international institutions and coercive regulations that deliver what justice requires. My aim here is to assess the permissibility of nudging and the regulation of other behavioral influences in light of institutional deficiencies and the socioeconomic conditions contributing to them and to partial compliance. This does not entail that this realism about institutions is complacent, in the sense that it hardly at all deviates from institutions as they are (Estlund 2014). With the exploration of patterns of behavior at low capacity, the development of techniques of steering it, and analyses regarding how these techniques may be accommodated into the set of practicable institutional regulations, we can adopt a moderately aspirational realism in our non-ideal theorizing.

Finally, my non-ideal approach is transitional as opposed to end-state, meaning that it is meant to "guide action in our current circumstances" (Stemplowska and Swift 2012, 385). Namely, it does not follow from the permissibility of a particular behavioral influence that it could not or should not later be replaced by some improved policy able to deliver full justice. This is, however, a cautious optimism. I only leave it open that the evolution of our liberal democratic institutions might be able to deliver policies that are more effective and more just than the ones I defend here. Whether such policies, and in turn, such institutions are achievable will be resolved in due time. All that my non-ideal account seeks to do is normatively assess whether behavioral techniques can narrow the gap between partial and full compliance and upgrade our institutional assets, thus improving the conditions in which individuals can act autonomously and fulfill their moral duties.

1.2.4. ***Anti-perfectionism***

Finally, we must determine the species of liberalism that lies at the base of my argument in favor of nudging and regulating non-reflective influences, with regard to whether it allows promoting the good. Liberal perfectionism holds that it is "at least sometimes permissible for a liberal state to promote or discourage particular activities, or ways of life on grounds relating to their inherent or intrinsic value, or on the basis of other metaphysical claims" (Quong 2010, 27). Liberal anti-perfectionists, on the other hand, claim that it is no business of the state to promote policies derived from comprehensive theories of the good. If justifications of government nudging were grounded on some objective conception of what constitutes well-being or flourishing, or would hold to flourishing itself in a comprehensive way, then such justifications would be regarded impermissible among anti-perfectionist liberals.

Justifications of nudging, those following in the footsteps of Thaler and Sunstein in that they aim to benefit individuals 'by their own lights', would not prima facie be perfectionist because such justifications do not appeal to or favor any specific metaphysical claim about the good. However, it might be difficult to square a narrow conception of nudging with anti-perfectionism. First, even the subtitle of Thaler and Sunstein's book makes clear references to improvements in the area of health and wealth; if such values are pursued for their own sake, then policies that promote them would more closely fit a perfectionist justification than an anti-perfectionist one. Second, a nudging government's act of protecting autonomy on any given conception, although sensitive to diversity in preferences and values (‘by their own lights’), could still be perfectionist if grounded in a comprehensive view of autonomous flourishing. Finally, even if nudging governments aim to uphold anti-perfectionist principles, many of their actions will set up choice environments that will inevitably and predictably favor certain options against others. As Sunstein argues, much of government nudging will come by way of designing websites, or setting up frameworks for contract, property and tort law (2019, 21). The last point seems particularly worrisome for hopes of grounding an anti-perfectionist case for nudging in a narrow sense.

Still, I will try to offer an anti-perfectionist case for nudging and regulating non-reflective aspects of choice environments. This case will observe many disruptive influences that stifle the use of moral powers by liberal citizens, and will argue for nudging as a means of overcoming such influences, and thus *enabling* moral powers. Nudges will often do more than overcoming influences – they will aid individuals in pursuing their autonomous goals. In order to be truly anti-perfectionist, a case for nudging will have to employ safeguards making sure that dissenters of policies favoring a particular comprehensive view will be minimally burdened when they wish to go the other way. If my anti-perfectionist case for nudging and regulating influences more broadly is convincing, then such a view has the added value of having a truly liberal signature. This is because it can – obviously – appeal to anti-perfectionists, but also to perfectionists, who are not principally opposed to anti-perfectionist justifications.[28]

As an appendix to presenting the non-ideal and anti-perfectionist credentials of my work, let me say a few related words here about how the utilization of nudges morally compares to coercive government measures in my framework. In short, this depends on whether nudges are used for self-regarding or other-regarding purposes. With regard to the former, I will show

[28] That nudge policies can be made compatible with anti-perfectionist principles is suggested by the fact that Rawls himself, a committed anti-perfectionist, believes that irrationality-overcoming policies that promote the ends of individuals are permissible in the original position, as I showed in 1.2.1.

in Chapter 3 that nudging, with the principle of transparency in place, the purpose of which is to enable dissenters to circumvent their influence (and preserve the anti-perfectionist character of my position), can help autonomous pursuits. In that regard, nudges have a practical edge over coercive regulation, which will in most cases be uniform, and thus, less sensitive to autonomous pursuits. For instance, banning cigarettes is certainly less sensitive to differences in conceptions of the good than a nudge against smoking. On the other hand, nudging for other-regarding purposes, such as the alleviation of grave harm, will not be considered morally special compared to coercive regulation. My endorsement of such nudges will be advocated in the context of institutional shortcomings in producing effective regulation and the inability of individuals to abide by moral and/or legal rules.

1.3. Summary

In **Chapter 2**, I discuss how the appreciation of behavioral facts should complement discussions about personal autonomy, in order to determine what kind of autonomy liberal democracies must respect. I engage with the standard autonomy-related objections to nudging, and show that they prove too much; accepting them leads to the conclusion that we, for the most part, lead non-autonomous lives. Laying out the empirical foundations for the pervasiveness of non-reflective influences, which are an unavoidable aspect of our psychological lives, helps me to establish the foundations of an account of personal autonomy sensitive to behavioral facts – resource-based autonomy – which will be used to test the permissibility of various behavioral techniques and influences more broadly. Finally, I explore how this new account of autonomy fares against the inevitability of behavioral influences.

In **Chapter 3**, I turn to the transparency of nudge techniques that would respect the behaviorally updated account of autonomy. I explicate the notion of watchfulness within the broader principle of *in principle* token interference transparency of nudging proposed by Luc Bovens (2009). Bovens's *in principle* transparency requires that citizens are able to see the nudge if they are *watchful* (ibid., 216-217). I conceptualize watchfulness as an opportunity that citizens are able to seize with certain institutional provisions in place – a rudimentary education on cognitive heuristics, a disclosure of aims deliberated on with other citizens, a registry of nudges in usage, and expert whistleblowers and consultant. The principle allows citizens to navigate the world of nudges in order to effectively pursue their autonomous goals.

In **Chapter 4**, I take a step back to acknowledge that if behavioral influences can negatively affect personal autonomy, nudges by government, even without their new liberal and democratic credentials, pose a much lesser threat to autonomy than the host of influences utilized by marketers. I analyze the kinds of influences found in markets, and argue that they are more threatening by virtue of sheer numbers and a lack of any normative principle that would prevent the sinister exploitation of our cognitive heuristics. I argue that our normative commitments of behavioral enhancement should spill over to our considerations of mitigating the bad effects of market practices. The means of achieving this in practice could take the form of advertisement-free public zones, bans on commercials posing as news, opt-out Internet adblocking, and stronger requirements about advertising content.

Chapter 5 turns to other-regarding considerations for nudging. Influencing individuals into other-regarding behavior, I claim, is driven not only by enormous social benefits, but by the fact that their non-compliant behavior often results from pervasive influences dragging them in the wrong direction. I defend so-called moral nudges in cases of moral certainty, even without the democratic safeguards explicated in Chapter 3, and hint at special cases of moral uncertainty, one of which is discussed in Chapter 6. However, a worry arises that moral nudges – not curtailed by the principle of watchfulness – undermine one of the moral powers – to understand, apply, and act from a public conception of justice. While such worries cannot be fully dispelled, I claim that they are somewhat exaggerated, given that the utilization of market nudges makes compliance with moral reasons no worse than it would have been in its absence.

Finally, **Chapter 6** gives a practical example of nudges that are controversial due to a clash of self-regarding and other-regarding considerations (a case of value uncertainty). Specifically, I look into existing nudge techniques that encourage charitable giving. The case of charitable giving reflects stalemates where it is obvious that there is a collective duty to be discharged, but also that we should allow citizens to opt out in moments of inconvenience. I discuss here the permissibility of using nudge techniques as the core of a collective endeavor (a nudge ethos) to discharge duties on people who disagree with the collective aim, or those who fail to act on their commitments. I also discuss whether nudges can be permissibly used in these contexts to attempt changes in personal attitudes, if such changes are feasible.