The Liberal Politics of Behavioral Enhancement

By

Viktor Ivanković

(Word count: 74893)

Submitted to

Central European University

Department of Political Science

In partial fulfillment of the requirements

For the degree of Doctor of Philosophy

Supervisor:

Andrés Moles

Budapest, Hungary

2019

Copyright notice

Viktor Ivanković

September 11th, 2019

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.

Acknowledgements

The opportunity to lay out the immense complexity of one's philosophical thought in a doctoral dissertation allows many to thrive. Not me. For me, facing its enormity and tying up a million threads made for a harrowing experience at best. Yet, the struggle makes the assistance I received to overcome this obstacle all the more precious. Firstly, I'd like to thank my supervisor Andrés Moles, for his patient and generous help and for his dedication to always give me useful and challenging comments. Not only has Andrés introduced me to a variety of philosophical questions, including the ethics of nudging, but has played a significant part in my growth as an academic philosopher. I would also like to thank Zoltàn Miklósi and Jànos Kis for their patient guidance in helping me shape the dissertation, as well as pointing out its strong and weak points.

I owe a great debt of gratitude to my friends Bart Engelen, with whom I co-authored the paper that later became Chapter 3, and Zlata Božac, with whom I developed and co-presented the ideas that later became Chapter 6. I was fortunate to receive tremendous help, academic and otherwise, from my wonderful friends and colleagues – Lovro Savić, Miles Maftean, Aleksandar Simić, Man-Kong Li, Megan Foster, Alfred Archer, Miklós Zala, Mihovil Lukić, Fay Niker, and Elettra Repetto. I also received fantastic comments on parts of the manuscript by Matthew Clayton, Luc Bovens, Andreas Schmidt, Chris Mills, Victor Tadros, and Ali Emre Benli. Some parts of the dissertation got scrapped along the way; be that as it may, many thanks to the people who helped me while I was still developing them – Fabienne Peter, Mat Coakley, and Jonathan Wolff. Finally, I am thankful to people who have helped me, apart from the dissertation itself, to improve my academic skillset and lay the foundations for my academic career – Tom Douglas, Elvio Baccarini, Luca Malatesti, and Neven Petrović.

I am grateful for the grants and financial boons that I have received: to the Tilburg Center for Logic, Ethics, and Philosophy for the visiting fellowship in 2016, to CEU for the Doctoral Research Support Grant in 2017 (allowing me to spend a semester at the University of Warwick), and to countless conference organizers. Without these, my many collaborations would have been impossible.

Last but not least, I would like to thank my nearest and dearest. To my loving family – my parents, my brother, and my sister – thank you for the unrelenting support, patience, and tolerance. To my partner Anamarija – thank you for being the best companion and supporter a person can ever wish for, for helping me snap out of my low points, and for all your love and kindness.

Table of Contents

Introduction **1**
1.1. What are 'nudges'? 7
1.1.1. Cognitive heuristics and biases 8
1.1.2. The elusiveness of nudges 12
1.1.3. The popularity of nudge units 17
1.2. Theoretical foundations 19
1.2.1. Political liberalism 19
1.2.2. Beyond the fixation on libertarian paternalism 23
1.2.2.1. Libertarianism 24
1.2.2.2. Paternalism 25
1.2.3. Non-ideal theory 28
1.2.4. Anti-perfectionism 31
1.3. Summary 33
Chapter 2: Autonomy in a Behavioral World **36**
2.1. Behavioral influences and the standard reflection-based view 38
2.2. Grounding the pervasiveness stipulation 44
2.3. Drawing up the contours of autonomy in a behavioral world 50
2.3.1. Autonomy and limited reflective resources 50
2.3.2. The limits of outsourcing self-government 60
2.4. The inevitability argument 64
2.4.1. Brute and intentional influences on autonomy 66
2.4.2. Acts and omissions 72
2.4.3. The advancement of behavioral science 77
Chapter 3: Transparent Nudges **80**
3.1. The worry, the responses, and why they do not convince 81
3.2. Transparency and the effectiveness of nudges 84
3.3. Transparency: a typology 87
3.4. Reducing the scope of inquiry 91
3.5. Watchfulness: importance and four conceptions 97
3.5.1. The minimalist conception 100

3.5.2. The maximalist conception ... 101
3.5.3. The whistleblower conception ... 102
3.5.4. The democratic conception ... 105
3.6. Watchfulness: normative demands ... 107
3.7. Objections ... 112
Chapter 4: Market Nudges ... 118
4.1. Normatively engaging with market nudges ... 120
4.2. Market nudges – understanding and distinctions ... 121
4.3. Detailing the autonomy argument against market nudges ... 128
4.4. Objections ... 135
4.4.1. White ... 136
4.4.2. Further objections ... 140
4.4.2.1. Demandingness ... 140
4.4.2.2. Conceivability ... 143
4.4.2.3. The appropriate sites of moral assessment and institutional regulation ... 146
4.5. Policy applications ... 150
Chapter 5: Moral Nudges ... 155
5.1. Behavioral influences on moral abidance ... 158
5.1.1. Detrimental influences ... 158
5.1.2. Moral belief and moral motivation ... 161
5.1.3. State-of-the-art moral nudges ... 163
5.2. Enforceable and non-enforceable duties ... 168
5.3. The *pro tanto* wrongness of moral nudges – considerations from the moral bioenhancement debate ... 176
5.3.1. Moral compliance ... 180
5.3.2. Value uniformity ... 185
5.4. Strengthening the case for moral nudging ... 189
5.4.1. The inevitability argument ... 189
5.4.2. Scarce resources ... 191
5.4.3. Personal autonomy ... 193
Chapter 6: Charity Nudges ... 195
6.1. The potential of charity nudges ... 196
6.1.1. Studies and heuristics ... 197

6.2. What is there at stake with charity nudges? 200
6.2.1. Charity nudges and personal autonomy 201
6.2.2. Biases against philanthropy 208
6.2.3. Duty of poverty alleviation 210
6.3. The nudge ethos 212
6.3.1. Objections 218
Conclusion 224
Bibliography 228

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

[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]).

legal regulation, claims Farrell, provokes contestation and protest, and sometimes civil disobedience. Bad nudges go unnoticed, or are merely sidestepped (ibid.).

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

[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.

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 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

[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.

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 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

[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).

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 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'

[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).

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 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

[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).

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. 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

[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).

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 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

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

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-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).

[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).

[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.

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 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.

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

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. However, Sunstein's

[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).

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),

[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.

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 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

[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).

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 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

[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.

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 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]

[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.

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-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

[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).

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*", 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 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]

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

[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

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 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

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.

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 full instead of partial. 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 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

[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.

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 consultants. 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 moral 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.

Chapter 2: Autonomy in a Behavioral World

In the Introduction, I asserted that nudging and behavioral influences more broadly might disturb our capacity to form, revise, and pursue our conceptions of the good life. Such considerations are elaborated in the philosophical debate on personal autonomy. Most objections to nudging, as I show here, are autonomy-related in one way or another. They either explicitly state that nudging undercuts reflective cognition, which is standardly thought to lie at the core of autonomous choices and pursuits, or implicitly by reference to autonomy-related concepts like manipulation, agency, and transparency. I argue in this chapter that addressing autonomy-related objections to nudging requires a prior normative undertaking – providing an account of autonomy that accommodates the pervasiveness of non-reflective behavioral influences. Only when our account of autonomy is updated with relevant facts from the behavioral sciences can we hope to assess the permissibility of nudge techniques. In this chapter, I draw up the first contours of such an account.

Here, I stipulate the pervasiveness of non-reflective behavioral influences (hereinafter, 'pervasiveness') as a factual aspect of our psychological lives. The stipulation is that a significant portion of human choices and actions are inevitably steered by non-reflective elements and operate at low cognitive capacity or fully automatically. The 'elements' here concern the workings of cognitive heuristics and other episodes of low-capacity operating (for instance, peripheral operations as opposed to focused action, or habits and trained skill as opposed to unfamiliar problem solving). The nudge literature helps to point to the empirical grounds for such a claim: a deeper appreciation of the impact of cognitive heuristics, which I have teased out in the Introduction, should strongly support pervasiveness. In other words, people should be expected, as part of their cognitive make-up, to consciously and non-consciously divert some of their actions

and decisions to automatic or low-capacity processing. For my purposes here, the stipulation should be assumed to hold even when individuals strongly favor handling their lives as reflectively as possible, and are set on activating their reflective capacities as much as they can. Differences in opinion are bound to remain concerning just *how pervasive* non-reflective influences and operations are. My stipulation establishes that, at a minimum, they are present to some degree in even the most reflectively led lives. Still, my reading allows a capacity for reflection that individuals can utilize and at least sometimes correctly identify in themselves as deliberation of the higher order.

I argue that an account of autonomy grounded on a deeper appreciation of behavioral facts strays from reflection-based accounts that have dominated moral and political philosophy over the last decades. My view stipulates the basic capacity for reflection sufficient for re-evaluating plans and considered preferences as a necessary and minimal condition for autonomy, but deviates from the standard reflection-based view, which is "not coherently applicable in a 'behavioral world'" (Schubert 2015, 3). The standard view cannot fully account for how the automatic and low-capacity processing can contribute to autonomous pursuits and complement scarce reflective resources. If we expect persons to make decisions and act automatically or at low capacity, due to their cognitive make-up, then having their reasoning bypassed in decision-making – a typical objection to nudges – will not itself suffice in rendering them non-autonomous. My hope in this chapter is that a serious consideration of automatic and low-capacity operations can help us draw a less abstract and more psychologically informed picture of personal autonomy.

I first present several autonomy-related objections from the nudge literature, explaining the background assumptions with which they are raised. Second, I add more empirical depth to the assumption of pervasiveness. Third, I give a first sketch of how these insights may contribute to a

more empirically developed account of personal autonomy. Finally, I discuss the inevitability of non-reflective influences, and analyze two ways in which non-reflective influences may come to undermine autonomy – in terms of their brute properties and in terms of the intentions of others.

2.1. Behavioral influences and the standard reflection-based view

The autonomy-related objections to nudging in the literature follow a legacy of what might be called reflection-based views of autonomy. Let's consider some of these objections in order to highlight the background assumptions with which they are raised. Daniel Hausman and Brynn Welch, for instance, object to nudging on the grounds that it diminishes "the extent to which [individuals] have control over their own evaluations and deliberations" (2010, 128). Similarly, Bovens states that "[w]hen we are subject to the mechanisms that are studied in 'the science of choice', then we are not fully in control of our actions" (2009, 209). The mechanisms employed by nudgers are rejected for effectively blocking the autonomy-related requirement "to act intentionally" (Beraldo 2017, 2), and are often considered outright "*manipulative* […] because [governments] deliberately circumvent people's rational reasoning and deliberating faculties, and instead seek to influence their choices through knowledge of the biases to which they are susceptible" (Grüne-Yanoff 2012, 636). Because nudges are thought to undermine people's deliberation about alternatives in decision-making, Hausman and Welch find nudging to be just as threatening to autonomy as coercion, if not more, given its covert form (2010, 130). Furthermore, if we establish that nudges "bypass/thwart/corrupt/distort the understanding, reasoning, and decisions of the individual", then it would be implied that nudges undermine important autonomy-based values in society, such as informed consent (Ploug and Holm 2015, 31).

These objections capture a commonsensical understanding of autonomy, typically tied to notions of self-government, self-legislation, and self-control (Feinberg 1986a, 28). I will focus particularly on 'self-government', standardly used as a synonym for personal autonomy in the literature. Those who care about people governing themselves are interested in a psychological ability of persons to be authors of their plans and actions. Satisfying these 'internalist' conditions of autonomy involves deliberating and acting on preferences with which a person identifies. The psychological character of such internalist conditions remains contested. I discuss two conditions here. On the first condition, a person acts authentically, and thus autonomously, if his second-order attitudes, or pro-attitudes, cohere with his first-order motivations (Frankfurt 1988; Dworkin 1981). While a first-order desire expresses what agents *want* (for instance, to eat chocolate), the second-order desire describes how agents relate to their first-order desire (whether they *want to want* to eat chocolate). We might call this the outcome-based condition of autonomy.[29] On the second internalist condition, which consists of two parts, the identification with lower-tier preferences is authentic, John Christman says, if the agent 1.) "reflects critically on a desire and, at the higher level, approves of having the desire" and 2.) "accepts the desire, value, or preference as part of her larger set of desires, beliefs, and principles, whether or not this is done for good reasons" (1988, 112). As far as authenticity is thus concerned, autonomous agents need not only reflectively ensure that the motivations that drive their actions are those they identify with, but that these fit within a greater structure of their psychological make-up, that is, within an ordering of the 'inner self' (Young 1980, 36). Call this the process-based condition of autonomy.

[29] The condition is the primary concern of coherentist accounts of autonomy. See Frankfurt (1988).

The exact kind of process that identification with first-order preferences requires is indeterminate, but will typically involve at least some reflective deliberation. Christman, for instance, states that a "person P is autonomous relative to some desire D if:

1. P was in a position to reflect upon the processes involved in the development of D;
2. P did not resist the development of D when attending to this process of development, or P *would not have* resisted that development had P attended to the process;
3. The lack of resistance to the development of D did not take place (or would not have) under the influence of factors that inhibit self-reflection (unless exposure to such factors are autonomously chosen, in which case that choice had to be made without such factors); and
4. The judgments involved in this self-reflection, plus the desire set that results, are minimally rational for P" (Christman 1991a, 347)

Christman believes that for preferences to be changed autonomously, we need to be able to distinguish between those processes that preserve autonomy, and those that do not. For him, it is preserved in a state of rational and reflective awareness from which changes in preferences come about (ibid., 348). This state of reflective awareness involves the capacity of an agent to "become aware of the beliefs and desires that move her to act" (1991b, 11, 17).

Non-reflective influences like nudges might undermine both the outcome-based and process-based conditions of internalism. First, non-reflective influences have the capacity to undermine the outcome of bringing first-order and second-order desires out of synchrony, by getting individuals to act on preferences with which they do not identify[30]; also, within their greater psychological structure, they may cause different preferences to become inconsistent with each other. Second, non-reflective influence, by its definition, could undermine the deliberative processes which characterize autonomous preference change, or the levels of awareness regarding

[30] It remains an empirical question whether non-reflective influences like nudges can also transform pro-attitudes directly, akin to some method of brainwashing. I say 'directly' because agents could change their pro-attitudes as an indirect result of acting on first-order preferences that were non-reflectively influenced. For instance, you are non-reflectively influenced to eat healthy and as a result of good consequences, as judged by yourself, you adopt a stronger pro-attitude towards healthy eating. Although your initial acts of healthy eating should, *ex hypothesi*, be rendered non-autonomous, your decision to change your pro-attitude on the basis of your reasoning would not seem to offend autonomy simply because the reasons for your decision have been highlighted by non-reflective processes. This should be empirically and normatively distinguished from processes that warp pro-attitudes themselves, if such indeed exist.

action-guiding motivations that autonomous action requires; as it is often stated in objections to nudges, non-reflective influences bypass reasoning in an autonomy-undermining way.

The externalist condition of autonomy, on the other hand, is concerned with whether the processes that characterize autonomy come from sources that could be called *our own*. Reflecting on our values and aligning our preferences with them "cannot be the whole story of autonomy" (Dworkin 1988, 18) because these processes depend "upon how [agents] came to possess values and desires that guide self-reflection, decision making, and the like: it depends […] on agents' *causal histories*" (Mele 1995, 146; emphasis in original; see also Christman 1991b). For a person's preference structure to truly be his own, the person needs to be *procedurally independent* when constructing it (Dworkin 1976, 25; 1988, 18; Young 1980, 36), meaning that his psychological processes must not be afflicted by a 'tainted source' (Mele 1995, 122) such as coercion, manipulation, indoctrination, or deception. Consider Alfred Mele's example of Ann:

> "Suppose that [Ann] *intentionally* voted Republican. Is that sufficient for her having autonomously voted Republican? No. Her intention to vote Republican might have issued from a posthypnotic suggestion. What Ann is missing in this scenario, to put it simply, is control over what she intends. […] Just imagine that the values of Ann's that played a decisive role in her deliberation were products of coercive brainwashing." (ibid., 12; emphasis in original)

Violation of the externalist condition occurs when the source of influence is tainted due to, for instance, the underhanded interference of others, as with manipulation, causing the governance of our own faculties to be overthrown and overtaken. We specifically resent our 'self' being replaced by a conscious 'other'. But we could imagine the autonomy-undermining sources to be tainted in different ways. For instance, we often conceive of the 'self' being eliminated from 'self-government' by an offset of severe mental illness, or persisting self-deception. Or, according to Lubomira Radoilska, "a person in the grips of compulsion does not act by choice. In this respect, her motives are not sufficiently up to her" (2012, 253). If certain influences on psychology are

tainted without them requiring a designer, then externalists should not only be worried about nudging, but about similar influences on behavior which are not designed by anyone. Otherwise, the externalist needs an account for why some non-reflective influences should be considered tainted, and some should not.[31]

Similarly, internalist conditions need to give us a non-arbitrary characterization of autonomy-preserving mental capacities and activities that leans on psychological findings and resists a commonsensical fixation on the psychological processes that we took to be at the core of autonomy. In the words of Jennifer Blumenthal-Barby and Aanand Naik, "cognitive science is problematizing this ideal with its empirical findings – our goals, desires and plans are less often the result of "rational" deliberative processes than we like to think" (2015, 45). Another possibility about why we take reflective deliberation to be at the core of autonomy is not so much commonsensical, but stereotypically philosophical. Dworkin believes that to make the reasoning capacities the measure of autonomy would constitute a "rather parochial view, one perhaps endemic to philosophers" (1976, 27). Similarly, Richard Double states that to philosophers, whose line of work "is strongly geared toward reflection, this requirement [of reflectiveness] cannot help but be terribly gratifying" (1992, 72-73). Following this, Double infers that we should resist setting

[31] On some externalist accounts, it is not even obvious that the usual suspects – coercion, manipulation, indoctrination – are sufficient for autonomy to be violated. DeGrazia (2012, 366) endorses a subjectivist account of the externalist condition, on which influences can violate autonomy if they are considered by the influenced person to be alienating. On this view, it is up to the influenced persons to decide, for their own particular cases, what constitutes alienating influence. Mele has a different picture of tainted sources in mind – brainwashing undermines autonomy regardless of whether the influenced person finds it objectionable (1995, 171). But in this case, externalism goes back to being concerned with setting the demarcating point between tainted and non-tainted sources. I do not offer a robust account of what makes an influence 'tainted' (or free of taint), and for the most part side with DeGrazia. It is the internalist conditions, however, that will be able to explain that, at one point, an influence (such as nudging) has become tainted. For instance, as I explain in 2.3.1. and 2.3.2., because non-reflective influences may marginalize reflective cognition beyond what is minimally required to sustain autonomy (regardless of what agents think), the addition of further non-reflective influences could be considered a tainting source.

up conditions for autonomy according to which "highly trained academics turn out to be vastly more autonomous than non-academics" (ibid., 73).

Still, it would be wrong to suggest that the psychologically updated view that I am to sketch here has no philosophical inspiration to draw on. Some philosophers have tried to distance us from reflection-based views. Sarah Buss claims that these 'super-agent' conceptions fail to show why we are not only accountable for behavior that stems from our intentional actions, but for less-than-reflective actions as well (for instance, actions resulting from akrasia). Contrary to reflection-based accounts, Buss claims that the autonomy of persons "depends on whether she can be identified with the direct, purely causal, nonrational influences on the formation of her intentions" (2012, 658). The person's identification would not reflect autonomy, Buss says, "if she forms her intention under the decisive nonrational influence of conditions that are elements or symptoms of human malfunctioning" (ibid., 660).

Double offers another alternative account, suggesting that we have to be less strict about the content of autonomous agency. Rather than fulfilling some objective criterion, namely engaging in reflective deliberation, Double believes that the characterization of autonomy should remain open to individual management styles (IMS), which allow personal takes on how to organize one's decisions and actions. For instance, persons may have very different opinions about how they want to balance out deliberation and "spontaneity" (Double 1992, 68-69). He claims that, for some, a rigid reflection-based account represents an antithesis of what autonomy truly entails:

> "The man-of-action, so the paradigm goes, shoots first and asks questions later. The true free spirit may not ask questions at all. And what about the millions of persons to whom reflecting on their lower-order psychological states is not only an infrequent occurrence, but is anathema to their individual management styles? For many persons, life is to be lived, not worried over." (ibid., 73)

Both accounts, as will be seen in 2.3., help in setting up an account of autonomy that is sensitive to the assumption of pervasiveness. However, neither account shows that we must entirely abandon the reflection-based account. Instead, I will argue in 2.3. that the assessment of whether a life is led autonomously depends, to a great extent, on whether individuals successfully distribute their reflective resources between projects and activities in ways they see fit. This will allow me to claim that an individual's activities in areas he considers relevant to his conception of the good and that he wishes to carry out reflectively will be evaluated with the help of at least some internalist and externalist conditions. We should also be able to say that forming and revising our life plans, as well as delegating the tasks to pursue them, should be achieved in conditions which do not obviously block procedural independence. I will thus look to preserve the core of the standard view for the formation and revision of human values and goals, but explore how the account is to be updated by behavioral insight. I follow Double's proposition that there can be vast differences in individual management styles, as well as different takes on how non-reflective action in different areas relates to the preservation of personal autonomy.

2.2. Grounding the pervasiveness stipulation

The stipulation that much of human reasoning and decision-making operates at low capacity or fully automatically, and that they are steered by non-reflective influences, is grounded here in dual-process and dual-system theories of cognition. These theories have been particularly popularized for wider audiences in recent years by Kahneman (2011), but have been developed, somewhat separately, for almost half a century by cognitive and social psychologists. Common to most renditions of these theories are the notions that one type or system of reasoning (or a

combined set of modules), System 1, is fast, effortless, sometimes fully automatic, older in evolutionary terms and with processes running in parallel to one another, while the other type or system, System 2, is slow, effortful, conscious, controlled, takes up plenty of energy, is more recent in evolutionary terms, and unable to sustain multiple processes running at once (Evans and Over 1996; Stanovich 2004).[32] These kinds are often called System 1 and System 2, which are terms that I will adopt here for reasons of simplicity. As Jonathan Evans notes, it is the second type of processes – System 2 – which marks differences between individuals, particularly those relating to general intelligence and cognitive capacity (2009, 33). The context and the character of the problem will deeply affect how it will be processed, although some variation should also be expected to obtain given differences between persons (Mercier and Sperber 2009, 149).

Keith Frankish and Jonathan Evans (2009, 3-4) suggest that dual-process theories have been appearing in rudimentary form for quite some time, and that the idea of mechanical and low-capacity processes taking up a great part of cognitive operations (resulting in the bulk of our behavior) dates back to Descartes and Leibniz. In recent philosophical work, Frankish (2009, 90-92) believes that an obvious example of a dual-process conception can be found in the distinction between personal and subpersonal levels of action in Daniel Dennett's *Content and Consciousness* (1969), where personal action is understood as caused by conscious beliefs and desires, while subpersonal action stems from subconscious mental processes. When we, in folk terms, refer to agents as intending things and doing them as a result, we standardly think of personal – System 2 – operating.

[32] As we can gather from differences in terms, there is some controversy concerning whether these processes should be conceived in *systemic* terms or not (e.g., Evans 2009, 34-35; Mercier and Sperber 2009). I do not have to pick one version of the theory or the other for my purposes here. The importance lies with adopting a fairly similar set of properties for the two types or systems, which I explicate here.

Since System 1 is intricately programmed to guide us towards our evolutionary goals – mainly survival and reproduction – while System 2 sometimes allows us to break away towards our particular and personal goals, it is System 2 that we more commonly identify with ourselves (Frankish 2009, 89). System 1 is powered up by the various heuristics I detailed in the Introduction and makes it possible for us to quickly and effortlessly respond to our choice environment in predictable and patterned ways. Yet, at the same time, it is more often perceived to be undermined by breakdowns that produce cognitive biases.[33] The emergence of System 2 is an evolutionary novelty in the human species that gives us flexibility to override the guidance of System 1 when we have individual reasons to do so. Since System 2 is more attuned to the needs of the individual as a separate organism, System 2 overriding System 1 processes is more likely to make that organism better off (Stanovich 2009, 55). According to Jonathan Evans and David Over:

> "The advantage of the dual-process system is that conscious reflective thought provides flexibility and foresight that the tacit system cannot, by its very nature, deliver. Most of the time our decision making is automatic and habitual, but it does not have to be that way […] consciousness gives us the possibility to deal with novelty and anticipate the future." (1996, 154)

As I explained in the previous section, much of our theorizing about autonomy is reflection-based. It primarily identifies the autonomous individual with his conscious responsiveness to reasons for action. Insofar, conceptions of autonomy typically reside with System 2, that is, the reflective aspect of human cognition. This is understandable, given the supervising role of System 2 over more automatic cognitive processes. This approach also corresponds with a folk psychological and pre-theoretical view of the human mind. However, we have good reasons to think it is an oversimplification. What are the matters of psychology that complicate the story?

[33] "The main reason", Evans explains, "that dual-process theories have been developed in the psychology of reasoning and decision making has been to account for cognitive biases and their apparent conflict with logical reasoning or normative, rule-based judgment and decision making" (2009, 40).

First, much of our behavior is produced with little or entirely without intervention of System 2 reasoning (Frankish 2009, 93). Playing a musical instrument, playing sports, or driving a car requires little reflective intervention; our success at performing these actions could in fact be undermined when reflective intervention does occur. A number of scientists are working to show that low-capacity (or, 'intuitive') decision-making is in fact highly effective and should not be associated too closely with "faulty reasoning"; as an evolutionary trait, System 1 processes reliably serve goals well most of the time (e.g., Myers 2002; Gigerenzer 2007). It is to System 1 processes that we owe the ability of habit formation and learned skill, which mark the transfer of reflective to non-reflective control (Stanovich 2009), thus unburdening our limited attention span. And although much of resulting action is viewed by agents to be part of autonomous pursuits of goals and values, it is often not responsive to reasons in the way abstract and decontextualized thinking of System 2 is.

Second, the psychological literature often conceives of System 2 as a 'capacity' and a 'limited resource'. A particularly popular idea is that of the 'cognitive miser' (Tversky and Kahneman 1974; Kahneman 2011), which entails that humans "engage the brain only when all else fails–and usually not even then" (Hull 2001, 37), and emphasizes "limited mental resources, reliance on irrelevant cues, and the difficulties of effortful correction" (Krueger and Funder 2004, 316-317) (from Stanovich 2009, 69). We should expect individuals to try to "ease the cognitive load and process less information" (ibid.) whenever possible. The cognitive miser engages in effortful cognitive activity "only when properly motivated–perhaps because subpersonal reasoning has failed to deliver a satisfactory response, or because we have reason to take special care" (Frankish 2009, 93). If humans are indeed cognitive misers, then they have a natural psychological tendency towards operating at low capacity, thus reinforcing the pervasiveness stipulation, and the

processes which we traditionally identify as lying at the core of autonomy will often be utilized only when all other processes have failed. My account of autonomy puts special emphasis on incorporating the idea of limited reflective resources.

Third, individuals often fail to realize that their actions were not reflectively responsive to reasons, but were in fact driven by System 1 processes. Frankish calls this kind of reasoning 'confabulatory', "serving merely to rationalize intuitive responses generated by nonconscious processes" (2009: 97). He believes that there is good evidence to believe a good deal of our reasoning is confabulatory (Gazzaniga 1998; Wegner 2002), but not all of it. System 2 is, in fact, also susceptible to cognitive bias, and often misses out on non-reflective triggers as well as their effects, giving us a false impression that it was our reflective responsiveness to reasons that caused the action.[34]

Consider that at least some of the forming and revising of our conception of the good, even if guided by responsiveness to reasons, will undoubtedly be influenced by cognitive bias. A common factor in setting up our future goals is the optimism bias – expectations of positive outcomes which are not supported by much evidence.[35] Yet, we might often believe that some life plans, even if somewhat irrational and unsupported by evidence, are internalized and integrated into the value structure of individuals in a way that should not be interfered when these inform them how to pursue their conception of the good life. We might believe that they are sometimes more expressive of our autonomy than some results of reflective processes.[36]

[34] Consider, for instance, the tendency of humans to underestimate the extent to which their behavior is susceptible to social conformity, as opposed to the susceptibility of those around them (Pronin et al. 2007).
[35] See, for example, Jefferson et al. (2017).
[36] I allow that non-reflective influences are compatible to an extent with forming one's conception of the good, but the reader will excuse me for not saying exactly *how much*. I leave this to future considerations and empirical updates. I will assume, for the time being, that the usual suspects, like brainwashing and oppressive upbringing, are not compatible with independently forming one's conception of the good.

Hence, the intricate balance between System 1 and System 2 processes will paint a complex picture of personal autonomy. It is the contours of this picture that I am trying to sketch in this chapter and throughout the dissertation. The challenge is to determine how the balance is to be struck between System 1 and limited reflective resources of System 2. Certain usages of both kinds of processes, in the grand scheme of one's cognitive architecture, can be supportive as well as undermining of personal autonomy. Blumenthal-Barby, for instance, convincingly argues that certain operations of System 1 may be thought to undermine personal autonomy. While making medical decisions, cognitive biases and heuristics (like the optimism bias, impact bias, and framing effects), even if they were not deliberately triggered by anyone in particular, can seriously undermine a person's understanding of the circumstances, available actions, and consequences. Heuristics and biases could bypass the intentions of agents, clash and override them, or hinder them in forming plans and intentions (for instance, a default in health care is a significant predictor of medical decisions) (Blumenthal-Barby 2016, 9-11).[37]

Some would, I venture, still believe that the behavioral terrain that I am trying to lay out for considerations of personal autonomy is too distant from commonsensical views. An agency constrained by pervasive non-reflective influences that I described here is a far cry from the ideal of the reasoning-responsive individual. If the body of research in behavioral science correctly ascribes non-reflective elements to a vast area of human action previously considered fully reflective, then autonomy might have simply been an illusion that we should now discard. And they would be right that we can at least conceive of the point at which low-capacity and automatic action effectively blocks individuals from being truly self-governing. Still, even if we have to give up on a pure ideal of autonomy, we might still be able to differentiate between modes of cognition

[37] For a more extensive list of examples and reasoning, see Blumenthal-Barby's referenced paper.

that block autonomy altogether, and those that do not. For instance, mental disorders seem to involve symptoms that make it difficult, and in some cases impossible for individuals to preserve autonomy, whereas cognitive heuristics, which often make it easier for individuals to attain their goals, have a more ambivalent relationship with autonomy. Skeptics about the compatibility between autonomy and pervasiveness should at the very least be convinced that certain cases of balancing automatic and reflective processing resemble autonomy more than other cases. They should allow that a *lesser* autonomy – call it 'autonomy*' – can still be discerned from cases in which autonomy is completely absent in an agent. These skeptics can make sense of my claims here by reading my references to 'autonomy' as references to 'autonomy*'.

2.3. Drawing up the contours of autonomy in a behavioral world

Previous sections have set the stage for a behaviorally updated account of personal autonomy. This account will by no means be complete here. Rather, it will hint at how autonomy should be contained in the face of behavioral facts, and which new considerations these facts bring to bear. The section is divided into two subsections. The first sets up the account by continuing the discussion on the relation between non-reflective influences and autonomy, bearing in mind limited reflective resources. The second discusses the extent to which self-government can be 'outsourced', while still retaining a true sense of individuals governing themselves.

2.3.1. ***Autonomy and limited reflective resources***

Humans have limited reflective resources (Schubert 2015, 3). This suggestion is not very outlandish, and, as I have pointed out in 2.2., has become fairly commonplace among social

psychologists, cognitive psychologists, and behavioral economists.[38] We find it difficult to be equally reflective in all of our activities, and even if we could, many of us would likely find it undesirable. Permanently operating at high capacity can also, in many cases, decrease the success of our actions (when it replaces habits and trained skill), cause cognitive exhaustion, or, due to a limited attention span, lead us to undermine our central projects for the sake of activities we find less important.[39] I will claim here that an assessment of autonomy sensitive to behavioral facts should observe how, as part of their conceptions of the good, individuals prefer to invest their reflective resources. Call this resource-based autonomy. I argue that different conditions should be applied to assessing whether decisions are autonomous depending on how individuals cognitively organize their activities. For some decisions, typically those central to a conception of the good, at least some internalist and externalist conditions from 2.1. will apply because individuals willingly operate reflectively. For other activities, some of which individuals consider trivial or would simply see them through without giving them much thought, stringent reflection-based conditions should be rejected. A more general assessment of whether individuals are governing themselves checks if they manage to organize their reflective and non-reflective lives according to their plans and goals.

Yet, simply because individuals want to carry out certain actions at low capacity does not mean that they are indifferent to them. We may choose to carry out certain central projects at low

[38] I expand on this in Chapter 3, where I discuss the concept of limited cognitive bandwidth, discussed at length by Mullainathan and Shafir (2013).

[39] Sandel believes Rawls himself anticipated the constraint of limited reflection. For Rawls, it is our life plans that stem from "careful reflection in which the agent reviewed, in the light of all the relevant facts, what it would be like to carry out these plans and thereby ascertained the course of action that would best realize his more fundamental desires" (1999a, 366). Sandel believes a limited scope for reflection is implied in this quote, since the outcomes of it seem to be restricted to choosing plans against alternatives and weighing the intensities of desires (1998, 159). But while Rawls's quote indeed seems to imply that there is a centrally appropriate way of focusing reflection, it is not quite certain from whence Sandel draws the conclusion that focused reflection is *restricted* in Rawls to central plans, and, thus, that the Rawlsian account regards reflection to be limited in scope.

capacity because we believe this will produce better results overall, or because some of our goals are so deep-seated that reflectively pursuing them will bear no impact on future revisions. Or, even if we do consider matters trivial, we may still have a preference regarding how to act at low capacity. Imagine, for instance, that the employees at a cafeteria do not consider dishwashing an activity relevant to their central values and pursuits. But while they do not wish to invest reflective resources into it, they may still prefer to wash so that no dish ends up broken. The employees could organize their faculties and environment to minimize dish-breaking, or have their faculties and environment organized for them, at no apparent loss to autonomy.

That non-reflective influences can complement autonomy in this way is acknowledged even by classical autonomy theorists. For instance, Christman claims that resisting a purely reflective account of autonomy is due to the fact "that people often subject themselves, in ways and under conditions that manifest autonomy, to factors and influences that severely undercut their reflective capacities. This is done, however, to accomplish things (or states of mind) that cannot otherwise be achieved" (1991b, 20; see also Christman 1991a, 347). Willingly operating at low-capacity in some area thus seems autonomy-preserving, even if the arrangement of non-reflective influences is left to some third party.[40] This grounds the case of some authors that nudges, although manipulative, are consistent with autonomy if they are consented to (Wilkinson 2013, 353-354).[41] This allows autonomous individuals to outsource some of their self-government, which I will further discuss later in this chapter, and in Chapter 3.

Yet, in some areas of life, choice environments will not be organized, nor will individuals give consent to being non-reflectively influenced. Some authors think that because there is a vast

[40] Both of these claims are not unconditionally correct. I tackle the exceptional circumstances in 2.3.2.
[41] See also Schubert (2015, 8) and Lades (2014); the latter believes that nudges can be conceived as 'behavioral self-commitment'.

range of such areas, "every human agent needs to rely, in one way or another, on her surrounding choice architecture in order to economize on scarce mental resources" (Schubert 2015, 6-7). An even tighter connection between autonomy and the economizing of mental resources is drawn by Sunstein, who states that:

> "[e]very hour of every day, choices are implicitly made for us, by both private and public institutions, and we are both better off and more autonomous as a result. If we had to make all decisions that are relevant to us, without the assistance of helpful choice architecture, we would be far less free […] If we had to make far more decisions, our autonomy would be badly compromised, because we would be unable to focus on what concerns us." (2015b, 137)

This passage suggests that non-reflective influences promote autonomy even when they are not consented to. Although this claim can be true in some cases, we should be much more cautious here. Many non-reflective influences, including some nudges, will have the capacity to violate autonomy in cases where individuals aim to process decisions as reflectively as possible. Nudges may bypass these processes or blur the preferences that move individuals to act. Additionally, in such cases, a nudge which is particularly difficult to resist usurps the capacities of individuals to govern themselves, when they want to do so reflectively. But let's first concentrate on cases where non-reflective influences which are not consented to can preserve autonomy.

Take two persons, Amy and Holden. Both invest their reflective resources in ways that help them realize their central autonomous goals, in line with internalist and externalist conditions. All other projects, secondary to their conceptions of the good, are handled by Amy and Holden in an unorganized low-capacity manner or fully automatically. Now assume that, as a matter of interaction between Holden's cognitive heuristics and the choice environment, he behaves in ways that are contrary to his second-order desires, secondary though these activities may be. Contrary to his wishes, Holden conforms with many social norms. He makes choices at the physician that he later regrets. He falls for ads and buys things he thinks he does not need. For this reason, Holden

either has to invest some of his reflective resources into these activities, thus potentially undermining his central pursuits due to limited reflective resources, or allow a break between first-order and second-order desires for the secondary activities. Amy, on the other hand, behaves optimally when operating at low capacity. It stands to reason, I think, that Amy enjoys better conditions for an autonomous life than Holden. This should imply that non-reflective influences could promote autonomy, even when they are not consented to.

Some might suggest that Amy's autonomy is not exactly "promoted" by a suitable interaction between her cognitive heuristics and the choice environment. Autonomy, after all, is realized by how individuals *govern themselves*, which is, the argument would suggest, not the function of low-capacity processing. Autonomy* (see 2.2.), on the other hand, at least in areas where individuals would allow themselves to act non-reflectively, might be promoted by achieving mere coherence between first-order and second-order desires, or undermined when they become disjointed, or so I will claim. One possibility is that the preservation of autonomy in the non-reflective realm can be expressed in a counterfactual: an individual acts autonomously if he acts in a way he would reflectively endorse had he chosen to invest available reflective resources into the process preceding the action.[42] Similar to Buss's account (2012), these would be non-reflective causes with which the individual can identify.

The objectors might still insist that a simple coherence test for non-reflective behavior is not an obvious normative solution for our modified autonomy account. But I do not have to settle

[42] This counterfactual condition is similar to Christman's second condition of autonomy from 2.1., which states that the individual has an autonomous preference if he would not have resisted its formation had he attended to it (1991a, 347). One relevant difference is that Christman's condition makes no mention of the availability of reflective resources. Christman's condition would be satisfied, it seems, if his agent could have attended to *all* of the processes of preference formation. But this implies a virtually unlimited pool of reflective resources, which is psychologically unfeasible. The counterfactual must incorporate a reflective resource constraint, so that the individual could have attended to the preference formation if he had available reflective resources or if he would have accepted not attending to other preference formations, thus freeing up the reflective resources required.

this issue here. I could safely suggest that Holden's autonomy is undermined, even if Amy's autonomy is not promoted, by virtue of the brute facts of non-reflective influences that steer their actions (more on this in 2.4.), and by virtue of Holden needing to spend more reflective resources on trivial matters. This is enough to claim that Amy is more autonomous than Holden.

Non-reflective influences, thus, have an ambivalent relationship with autonomy – at a minimum, they could promote it when influences are consented to or identified with; conversely, they could undermine it in a range of cases. Whether they will promote or violate autonomy will depend on how individuals prefer to manage their decision-making. To borrow from Double (1992), I suggest that decision-making styles will vastly differ among individuals. Some opt to lead more reflective lives, while others prefer a more carefree routine and purposefully leave much of their decision-making to low-capacity operating. Variations in capacities and management styles should be compatible with a resource-based autonomy account. However, there will be some objective limitations regarding the kinds of management styles that preserve autonomy. I mention two here.

Some people will use reflective resources in excess. This gives rise to the *overthinking objection*, which comes in two versions. First, aiming to cover as many areas of decision-making, individuals could, at some point, compromise the accomplishment of their goals due to limited reflective resources. Using reflective capacities in excess of personal limitations causes fatigue or makes individuals distribute their resources in ways that distract them from projects that matter to them. Second, in certain areas of decision-making, people expend reflective resources far beyond a point of usefulness, and may even fall into a state of decisional paralysis. In the words of Schaefer et al., such individuals "could have improved autonomy by *reducing* certain deliberative inclinations" (2014, 128; emphasis in original). Schaefer et al. hint that autonomy will rarely be

achieved in a state of extreme reflective exertion: "the proper solution is not to simply inhibit deliberation, but rather improve people's ability to calculate when deliberation is important to achieving their goals and values, and at which point it should be foregone" (ibid.).[43]

Other people will hardly use reflective resources at all. This gives rise to the *autopilot objection*, according to which individuals are non-autonomous if the better part of their lives is a mere product of non-reflective influences. Externalists could still be satisfied with an autopilot management style. As Christman says: "What must *not* be true […] is that the person's judgment and evaluations of her states are *all* the product of these [non-reflective] influences. At some point, the person had to have undertaken the exposure to these factors while not under the influence of such factors for her autonomy to be preserved" (1991b, 21; emphasis in original). For externalists, then, it would be sufficient that an autopilot management style was consented to under conditions of procedural independence. The worry, then, seems to be mainly internalist. It would state that forming, revising, and pursuing a conception of the good is hardly conceivable without a minimum of reflectiveness. Of course, this begs the question regarding what the minimum of reflectiveness is. My account requires individuals to at least endorse the management style, to reflect on whether their discernible goals and values are supported by their actions on autopilot, and to be able to revise both their goals and the management style.

It is possible that other kinds of management styles will be revealed by cognitive science to be incompatible with self-government in this updated sense. But I think a range of distinct management styles will remain compatible with resource-based autonomy. My account requires

[43] On a related note, this means that certain policy measures that I have mentioned in the Introduction, namely choice architectures which do not trigger heuristics and which I have characterized as non-threatening to autonomy, could in some rare cases undermine autonomy by further inducing the reflection of overthinkers. I believe, however, that the rarity of such cases allows me to cast this marginal issue aside here.

merely that individuals endorse a rough balance of cognitive modes that they see as fitting with the attainment of their goals. Let me explicate the view further by responding to two worries that may arise here.

The first worry states that individuals could have false beliefs when they constitute their management style. On one hand, false beliefs could concern themselves – they could be way off regarding how much reflective resources they have at their disposal or whether some management style serves their purposes. On the other, they could have false beliefs about behavioral facts that substantiate the pervasiveness stipulation. The 'self-made man', for instance, often grossly overestimates the extent to which outcomes were up to his reflective control.[44]

In classical discussions on autonomy, some views hold that the preservation of autonomy must depend on having certain true beliefs. In Christman's words, some views raise "the requirement that the beliefs upon which […] desires rest are based on an (objectively speaking) adequate degree of evidence" (1991b, 14). This would render the self-made man non-autonomous on the basis that his beliefs about his reflective control falsely project how much he truly governs himself. However, it seems to me that this requirement may undermine autonomy much more than it does autonomy*. Autonomy* should permit as a standard part of the human condition that many individuals have a plethora of false beliefs about their behavioral selves and the behavioral facts that guide their actions. If this were not the case, autonomy would only be possessed by a select few.[45] My account should allow autonomy to be compatible with at least some false beliefs. Still, false beliefs could undermine autonomy *indirectly*. For instance, false beliefs often lead to

[44] See, for example, Thompson (1999) for illusions of control, and Kahneman (2011, Chapter 20) for illusions of validity.

[45] Those satisfied with an elitist view of autonomy need not consider this a problem. But such a view, it seems to me, is more prone to resisting psychological updates, than incorporating them.

adopting management styles that are deeply unrealistic or unsuitable, and could end up undermining the pursuits of goals and values, both in terms of their success and the modes in which agents would see them achieved.[46]

The second worry is that structuring elaborate management styles to adequately balance out our cognitive modes still seems like a stringent requirement for being autonomous. The fear is that this could re-direct autonomy back to cumbersome reflection required to organize our management styles, and put it out of reach for a significant majority. This is analogous to a worry Christman raises for internalists, which concerns the demandingness of establishing the internal consistency of beliefs and desires. Christman's solution, I believe, hints at how we could resolve our problem here as well. He states that an autonomous person must not be "guided by *manifestly inconsistent* desires and beliefs", where 'manifestly inconsistent' refers to "preferences and beliefs that are in obvious conflict, ones which the agent could bring easily to consciousness and recognize as incompatible" (ibid., 15). Similarly, I believe autonomy should only require individuals not to exemplify behaviors which are manifestly inconsistent with their management styles in areas for which they have specifically adopted them. The way individuals conduct themselves has to be in obvious conflict with their decisions regarding their management styles.

Let me recapitulate the first contours of autonomy in a behavioral world. First, I have claimed that due to their limited reflective capacities, individuals should be expected to perform some of their tasks at low cognitive capacity, or automatically. Individuals will choose to invest their reflective resources in particular areas of decision-making, and by so doing, they will create

[46] In this sense, the man who obsesses about being self-made may fail the test of the overthinking objection. The attempt to maintain control over all aspects of our mental life, given the scarcity of reflective resources, may cause a life to be less autonomous than if the person aimed at far less control. This would make the control-driven conception of the self-made man self-defeating. I thank Andrés Moles for pointing this out to me.

their own management styles. Endorsing and acting in line with a particular management style, one which is not ruled out by empirical constraints, captures an important sense of governing oneself. I have argued that autonomy should permit a diversity of management styles, some of which may include expending reflective resources to their limits, hardly use them at all, invest them into central projects, carry out central projects with little reflective effort, etc. One management style could adopt nudging to help economize reflective resources. I discuss this possibility in Chapter 3.

Second, I have argued that the conditions of standard reflection-based accounts, namely internalist and externalist conditions, should be used to evaluate the autonomy of pursuing one's goals only when these actions are carried out reflectively.[47] Even this claim will require some qualification. In cases where preferences are inevitably context-dependent, non-reflective influences may impact choice no matter how much individuals would want to avoid them. Such influences have the capacity to undermine autonomous action, yet they may be less threatening than the influences they merely substitute. This is discussed at length in 2.4. With pervasiveness in place, an updated account of autonomy (or autonomy*) acknowledges that reflective domains of action often cannot avoid non-reflective influences. As for standards of assessing the autonomy of low-capacity actions, I have not provided a strong claim regarding which standard we should take. One part of the standard might be to adopt a counterfactual test, asking whether the preferences that drive an individual into action at low-capacity would be endorsed had he chosen to invest available reflective resources. Another might be to deny that non-reflective influences

[47] Notice that this judgment is about the pursuing of one's good, not forming a conception of it. I work with the assumption that individuals at this stage, although influenced by non-reflective elements, have already formed a conception of the good under conditions that would not be deemed oppressive, e.g. under threat of violence and punishment, or as a result of brainwashing. As I mentioned before, I leave the matter of exactly how much non-reflective influence undermines the forming of one's conception of the good for another time.

promote autonomy in a true sense given that they are not obviously part of self-government, but allow that they undermine autonomy if they drive action that is at odds with values and goals.

2.3.2. ***The limits of outsourcing self-government***

As I mentioned, some authors have proposed that nudging be conceived as a form of pre-commitment that citizens can consent to and preserve their autonomy. I will pursue a similar strategy in Chapter 3. However, nudging is a blunt instrument. If not personalized, it is hardly sensitive to the diversity of management styles and deep-seated values and preferences. For some, nudges will simplify making choices they have difficulties with and help them economize reflective resources. For others, nudges will divert action from their values and preferences and disrupt management styles. On my account, nudges are non-threatening to the autonomy of the former, but not of the latter. Granted, policy makers usually attempt to influence individuals in areas where their goals largely converge. Still, such policies have a tendency to ignore the preferences of the dissenting minority. In Chapter 3, I will devise an account of transparency that alleviates the burdens of minority members to circumvent influences with which they disagree, in order to protect their autonomy.

Before that, some puzzles remain to be solved about the character of resource-based autonomy. Take the following. Even when influences are fully consented to, or adopted, our intuitions that individuals are truly self-governing may slightly diverge depending on whether their environments are designed by others, or designed by themselves. Consider an individual, call him Bart, who subjects himself to a range of influences that help him achieve his goals, while investing reflective capacities minimally. Imagine that Bart has personally designed and calibrated the means that drive his behavior. Now imagine, instead, that Bart consents to the same means, which

are just as effective, but designed and calibrated by someone else. Intuitively, Bart seems more self-governing in the first case than in the second.

Mikhail Valdman makes the opposite claim. He argues that it is not obviously bad for agents to "outsource" their self-government, i.e. willingly surrender their decision-making to another agent, if the substituting decisions by the other agent were made to reflect their deepest goals and values (2010, 764). I want to explore here whether Valdman's argument can have implications on the badness of agents being exposed to environments designed to steer them in non-reflective ways, even if agents consent to them or identify with their direction. Surely, Bart can consent to or identify with some nudging, given that he can also completely give up on some instances of decision-making. For instance, he can employ experts to make financial, legal, or health-related decisions for him without giving up on autonomy; similarly, it should be allowed that he employs these experts to nudge him in directions that he finds desirable. But could Bart allow his decisions to be steered by experts in the majority of his life decisions, trivial or non-trivial, without endangering his autonomy? And is there added badness to having your life steered by others, as opposed to designing your own choice environment, or allowing yourself to be steered by non-reflective influences "in the wild"?

Let's take the second question first. There is surely an added threat to autonomy in being influenced by others compared to being exposed to scattered non-reflective influences. Having others non-reflectively influencing our decision-making is intuitively closer to the threat of having our wills supplanted, and, in a republican vein, to being dominated by the arbitrary interference of others (Pettit 1996). I do not mean to reject these suggestions, since influences designed by others may indeed exploit our vulnerabilities to make decisions against our benefit (this is elaborated on the case of markets in Chapter 4). However, the next subsection shows that 1.) threats to autonomy

by designed environments are not by default greater than threats by non-designed environments, and 2.) there is moral closeness between changing choice environments and allowing environments with foreseeable effects, sometimes making it impossible to avoid the influencer. I have also assumed here, following Valdman, that 'outsourcing self-government' consists in agents consenting to the influences or identifying with their directions, and influencers making bona fide attempts to honor their deepest values and goals.

Fears about outsourcing self-government may then stem from the belief that the influencer will be unable to track our goals and values across vast decision-making areas, or steer our choices towards goals in a manner that we find normatively fit, even when he has the best intentions. This is the reiteration of the worry from the Introduction that nudgers cannot be effective means paternalists. Chapter 3 tries to overcome this obstacle, so let us assume for the moment that these feasibility obstacles are eliminated. Bart leads his life just above the autopilot threshold, and consents to or approves of most, if not all of the non-reflective influences to which benevolent parties expose him. Is there anything autonomy-threatening about outsourcing self-government in this way?

Some suggestions might give us pause. The first might be that in the face of external influence, we should invest at least some of our reflective capacities into choices that are meaningful to us. This is derived from the *meaningfulness objection* that Valdman poses to himself. The objection states that while there might be nothing wrong about ceding decision-making authority in trivial matters, something might be lost if one gave up on those decisions that "bear significantly on the direction and shape of one's life" (2010, 774). Assume, for the case of non-reflective influence (instead of ceding decision-making wholesale), that agents willingly make meaningful decisions at high capacity less often than willingly operating at low capacity under the

influence of nudgers. Intuitively, it seems autonomy-threatening to exercise reflective control over meaningful decisions in our lives less often than allowing them to be steered by the reflective control of others. Valdman's response to this objection is that by conceding that decision-making lacks value (as it does in trivial matters), we require further explanation why we need to personally exercise decision-making that shapes our lives according to our wishes (ibid.). But such an explanation is not completely out of grasp. We may believe, as I stated, that we should exercise reflective control over decisions more often than having them steered by the reflective designs of others. How much we should in fact invest reflective resources to satisfy this condition will surely be a matter of disagreement. But the response clearly demonstrates that not finding it valuable to invest reflection in *some* of our decisions does not entail that we should find it lacking in value to invest reflection in *any of them*.

Second, investing reflective capacities might be an important part of developing an evaluative stance. Similar to my last point, Stephen Wall believes that "[s]ome deference to authority is compatible with autonomous agency, but total deference to others is not" (2016, 180). Wall grounds his case on the claim that our evaluative stance about goals and values is not fully developed, but develops (at least in part) as we make significant choices (ibid., 182). Granted, in the case of non-reflective influences, influenced agents still make choices themselves, although sometimes at low capacity. It should not be assumed that these choices bear no impact on the development of evaluative stances. Still, there might be a difference to how decision-making impacts the development of our evaluative stance depending on the cognitive mode in which we make them. The worry is further reinforced by the possibility that the deliberate design of others may have greater impact on one's evaluative stance, than the impact of one's invested reflective

resources. This argument is by no means conclusive for our normative considerations, but signals additional caution.[48]

It is due to these cautionary signals, as well as feasibility constraints, that my case for incorporating nudging into liberal democracy in Chapter 3 will recommend *humble and cautious nudge projects*. In order to make sure that nudges supplement, and not supplant autonomous agency in conditions of scarce reflective resources, the advocacy of nudging should make sure that the implemented influences do not overwhelm the citizenry.

2.4. The inevitability argument

An important objection to my behaviorally sensitive account of autonomy still remains hanging in the air. Agents may flag the areas in which they wish to invest reflective resources, but given pervasiveness, they can hardly avoid non-reflective influences, designed or otherwise, in these relevant domains of choice. Is their autonomy inevitably compromised? I turn my attention to Thaler and Sunstein's inevitability argument (IA), which gives us a taste of the hard cases regarding the permissibility of intentional influencing. It claims that nudging is permissible because it is inevitable, *tout court*. In its original form, the IA points to the importance of context dependency of preferences and choices. Imagine that for students, healthy eating is a domain for expending reflective resources. In Thaler and Sunstein's example, Carolyn, a manager of the school cafeteria, notices that arranging food items influences students in predictable ways. This is because students seem to choose items that are more visually salient to them. A number of

[48] For a more extensive and detailed elaboration of why making our own choices might be significant, which applies to my discussion about limiting the deference of arranging our choice environments to others, see Lecture 2 of Scanlon (1988).

strategies become open to Carolyn – she can influence students in order to maximize profits, she can try randomizing the layout, or she can make them reach for an apple instead of a chocolate bar to promote their health (Thaler and Sunstein 2008, 1-2). The IA is, arguably, "Thaler and Sunstein's most important argument for nudging" (Grill 2014, 142). Libertarian paternalists suggest that the normative lesson from the IA is the following: if choice contexts need to be arranged in *some* way, then it is best, in normative terms, to arrange them so they make agents better off than they would be in the face of alternative choice arrangements. Thaler and Sunstein contend that some form of nudging is inevitable, and that to oppose nudging is a literal non-starter (2008, 11).

Thaler and Sunstein seem to adopt a hidden premise, which is that nudging can take on unintentional forms. On their account, people can become nudgers even if they are blissfully ignorant about the character of non-reflective influences and have no intention to steer behavior in any direction. Of course, such a claim is dubious.[49] Even if we grant to Thaler and Sunstein that choice architecture is inevitable, we might contend, as some do (Grill 2014, 143), that it makes a great moral difference to autonomy whether choice contexts are products of "natural causes", or they are specifically designed to non-reflectively steer the behavior of unassuming agents. And since intentional choice architecture is certainly avoidable, Thaler and Sunstein's argument loses most of its normative force.[50]

[49] The claim may have been upheld by Thaler and Sunstein to convince policy makers of the supposed inevitability of their policy approach, and not just of nudging.

[50] In my conceptual framework, I will reject Thaler and Sunstein's notions of unintentional nudges, as this notion deeply confuses matters. In my usage, nudges will only be considered intentional interventions on behavior.

2.4.1. ***Brute and intentional influences on autonomy***

I argue that Thaler and Sunstein's untenable version of the IA distracts from the important lessons that it should have for considerations of personal autonomy. Consider a slightly different take on the IA by Thomas Douglas (unpublished manuscript), which he calls the 'mere substitution' defense of nudging. Douglas explains that the 'mere substitution' defense rests on the idea that nudging introduces no new kind of influence:

> "P1: For some intervention (or set of interventions) *x*, all of the influences that *x* exerts on individuals' choices are influences of a kind that would also have applied to those choices in the absence of *x*.
>
> P2: An intervention does not wrong its target(s) if all of the influences it exerts on their choices are influences of a kind that would also have applied to those choices in the absence of the intervention.
>
> C: Intervention (or set of interventions) *x* does not wrong its targets." (ibid.)

Whether the influence introduced by *x* qualifies as the same 'kind' of influence that applies in the absence of it will depend on what we regard to be the properties that demarcate one kind of influence from another. As I have mentioned earlier, the intentionality of influences could represent a relevant property in our assessment of whether the influences are indeed of the same kind. Specifically, intentionality seems to be a necessary condition for non-reflective influences to be considered manipulative, since 'manipulativeness' is attributed only to influences that are designed and where bypassing reflective deliberation is consciously undertaken by other agents.[51] But while intentionality is one autonomy-undermining property, we should not assume that the influences in the pre-designed choice context have no autonomy-related properties. Intentionality

[51] The property of intentionality will be further complicated later on, when I establish that if the choice architect can also reliably foresee the effects of untampered environments, there will be little moral difference between intended change and leaving environments as they are. Intentionality will no longer make an autonomy-relevant difference because acts and omissions will be inevitable to the choice architect. This will ground the evidence-based view, which says that the architect acts wrongly if he commands sufficient influence-related evidence, and allows the environment to steer us against our autonomous pursuits (regardless of whether this was the result of an act or omission). I thank Andrés Moles for helping me to formulate this view. The evidence-based view will be formulated further towards the end of the subsection.

itself is not inevitable, but other (potentially autonomy-undermining) properties of the influence are.

To clarify this point, Douglas distinguishes between *brute* and *intentional* threats to autonomy. The notion of brute threats points to the fact that non-designed and designed influences alike, say, in a food arrangement case, may threaten autonomy because reasoning is bypassed in much the same way. In other words, both intentional and other influences have the same brute (or physical) properties, which can themselves threaten autonomy, especially in conditions of pervasiveness. Intentional threats to autonomy, on the other hand, only obtain when bypassing occurs due to the intentional interferences of others. The way in which intentional influences wrong autonomy relates not only to the physical facts of the influences, but to how individuals treat each other, or perhaps what power relations obtain between them. Consider the following pair of cases for illustration. In the first case, while spring cleaning in a closet, someone slams the door behind you, and locks you in. In the second case, also while spring cleaning in the closet, a strong gale pushes the door behind you, and jams it (or by some strange machination, locks it). The first case includes an intentional threat to autonomy that the second case does not – in the sense that your lack of autonomy, being locked in the closet, results from the designs of others to keep you inside. However, in a strictly brute sense, there is no point of distinction between the cases as the physical facts of each leave you non-autonomous, by trapping you in the closet.

A possible suggestion could be that the addition of intentionality to non-reflective influences could on its own transform their brute properties. Douglas discusses three respects in which intentionality may have such effects in the cafeteria case. Intentionality may 1.) change the phenomenological properties of eye-level food salience, 2.) change behavioral direction compared to the influences they replace, or 3.) increase the strength of the influence.

First, the phenomenological properties for the target seem to be identical for intentional and non-intentional influences, given that they play on the same heuristic-triggering effects. It is at least not obvious that intentional influences have different phenomenological properties from their non-intentional counterparts. Second, while an intentional influence will often change behavioral direction, for instance, from unhealthy snacks to fruit, both influences share the brute property that they non-reflectively direct targets towards particular food items. Therefore, a change in direction introduces no new property that would run afoul of the mere substitution defense.

Finally, the best candidate for a change in brute properties is that the intentional influence could end up being much stronger than the unintentional influence it replaces. For the moment, I will cast aside Grüne-Yanoff's objection from the Introduction about our lack of reliable measure for behavioral influences, and assume that we could sometimes determine a difference in strength comparatively. It would then probably be true that employing influences intentionally would imply greater control of the designer over the strength of the influence. Still, whether the intentional influences will have greater strength will be a contingent matter. Assume that while guests of the cafeteria all want to follow balanced diets, they have strong cravings for unhealthy snacks, which become much stronger under the influence of an untampered cafeteria arrangement. Assume, further, that if Carolyn rearranges the cafeteria to stimulate healthy choices, the intentional influence would be much weaker than the strong influence it replaces. Such circumstances, exceptional to the expectation that intended influences will be stronger, should be considered commonplace. Note, also, that proponents of nudging standardly only advocate influences which do not effectively block off alternatives. It should therefore be safely assumed that the replacing influences will be fairly comparable to the influences in place. By 'comparable', I mean that their brute properties will not deviate to an extent that would render them different in kind.

This is not to deny that there is an added layer of autonomy-related considerations when influences are intentional, as I already mentioned several times throughout the chapter. We might think intentional influences could 'violate' autonomy, whereas non-intentional influence could merely 'undermine' or 'diminish' it. But if brute properties are also (inevitably) relevant for autonomy for at least a significant number of cases, then it is not certain why considerations of intentionality would always precede or outrank considerations of brute properties. I have strongly teased throughout the chapter that brute properties of behavioral influences have the capacity to undermine autonomy, absent third-party intentions. Notice that the examples I borrow from Blumenthal-Barby in 2.2. to substantiate the pervasiveness stipulation are influences that are not the product of design. It seems apparent that many, if not all of these influences would be relevant for autonomy on standard conceptions. The brute properties of non-reflective influences may compromise the internalist's process-based condition, by disrupting the capacities of individuals to reflectively deliberate on their pro-attitudes and practical decisions. They might also compromise the outcome-based condition, by diverting us from acting upon desires with which we identify in the second order.

On a conception more familiar to nudge theorists, brute properties seem to be sufficient for undermining control. Yashar Saghai claims that nudges are permissible if agents can easily resist their influence. Yet, his take on the easy resistibility condition can be explicated wholly in terms of brute properties of influences: 1.) the agent needs to have the capacity to bring the influence to his attention, 2.) the capacity to inhibit his triggered propensity to do as influenced, and 3.) he is

not subject to a further influence that would undermine the workings of the first two capacities (Saghai 2014, 487, 489).[52]

Should avoiding intentional threats to autonomy always be normatively prior to removing serious brute threats to autonomy? Let's go back to the cafeteria case. Recall that all the students attending Carolyn's cafeteria have strong considered judgments about the importance of healthy diets, but find it difficult to stop themselves from having the occasional sweet. Carolyn knows this. Assume also that Carolyn is aware that an untampered cafeteria arrangement, by virtue of its salience effects, creates practically irresistible urges for students to pick sweets instead of fruit. Carolyn also correctly predicts that rearranging the cafeteria would create a weaker, but sufficiently effective influence for most students to honor their diets; this influence will, like its predecessor, slip by some if not most of the students unnoticed. Finally, assume that Carolyn cares about respecting the students' autonomies. Is she allowed to rearrange the cafeteria, on autonomy grounds?[53]

On face value, I would be inclined to say 'yes'. As a result of Carolyn's intervention on the food arrangement, the reflections of students will be disrupted by a non-reflective influence that, *ex hypothesi*, is much less difficult for them to resist. Also, Carolyn's intervention on the

[52] I am not claiming that Saghai conceptualizes his conception of permissible nudging in strictly brute terms – he does not. This is because, I believe, his conceptual efforts relate to nudging, not to non-reflective influences in general. All I am suggesting is that his conditions for our choices to be controlled by us are explicated in strictly brute terms, and are applicable to assessing whether an influence undermines autonomy regardless of whether it was designed.

[53] It might be even more difficult, and realistic in the context of diverse aims that we encounter when we deliberate on whether to nudge, to try to determine whether Carolyn should be allowed to rearrange the cafeteria in *interpersonal cases*. Imagine that by rearranging the cafeteria, Carolyn makes it easier for Norman not to pick up a sweet (when he would rather pick up an apple), and thus to honor his diet. However, imagine that by rearranging the cafeteria, Carolyn introduces a burden for Beth, who is inclined to pick up the sweet. Is Carolyn allowed to rearrange the cafeteria? I believe the consequentalist response is adequate here. As I will show later, the ability to predict for Carolyn that either Norman or Beth will be burdened by some choice arrangement means that there will be little moral difference between *rearranging* the cafeteria and *leaving it untampered.* Therefore, Carolyn should be driven by considerations of who becomes *more burdened*. If it is Norman, then she should be allowed to rearrange the cafeteria. Carolyn will, however, have to introduce conditions of transparency for Beth, which I elaborate in the next chapter.

cafeteria arrangement will make it more likely that the students will be able to act on their pro-attitudes to follow a healthy diet. This is not to say that the added consideration of Carolyn's intentions is not an additional autonomy-undermining factor. Still, Carolyn's intervention might be more preserving of students' autonomies, all things considered, than the influences that undermine their autonomies by virtue of their brute properties.

Now, one objection, or alternative explanation of the case, would state that it is more obvious that brute influences relate to welfare losses than to autonomy concerns. On this reading, Carolyn's intervention would be interpreted as a sacrifice of some autonomy, but a reasonable one given the considerable welfare loss students would suffer had she not intervened. On this reading, only Carolyn's intervention is autonomy-threatening. But this interpretation would have to prove, contrary to my arguments here, why the brute properties of non-reflective influences are welfare-relevant, but not autonomy-relevant. It would have to establish why we should ignore that these brute properties may disrupt psychological processes so often claimed to be at the core of reflection-based accounts.

Say that we agree that brute considerations are autonomy-relevant. How do they weigh against considerations of manipulative intentions, when such considerations are at odds? Can we hope to balance them given our inability to measure the strength of non-reflective influences? And are these theoretical insights practically useful? These are difficult questions. I attempt to flesh out the technical and practical implications of these insights throughout the dissertation, and especially in Chapter 3. At this point, we can do little else but to suggest that when we weigh between non-reflective influences of comparable brute properties, intentions might be the morally relevant addition to rule against a particular influence.

Let me thus make my first alternative takeaways from the IA. First, while intentional influences are certainly not inevitable, brute influences are. Second, if brute properties matter for assessing autonomy, then it is not obvious that considerations of intentions always outweigh brute considerations in autonomy concerns.

2.4.2. ***Acts and omissions***

Let us now return to the standard objection against the IA, which states that *intentional* influences are not inevitable. Even if brute facts about an influence could outweigh considerations of intentionality, influences by design could still contain an added threat to autonomy compared to the unintended influences of similar brute properties. This is because intentional influences might be subverting our actions to the wills of others. I will now argue that even this suggestion is not as straightforward as it may seem. There will be many cases in which a choice architect will be unable to avoid either intentional intervention or allowing untampered environments with fully predictable effects. The latter, I will argue, is a "close moral cousin" to the former.

Notice that, in most of my examples, Carolyn is very knowledgeable about all things behavioral. Her knowledge is central to this part of the discussion. She effectively predicts not only the outcomes of her behavioral interventions, but also the outcome in which she does not intervene, or the outcomes of randomized arrangements. But if Carolyn can accurately predict the effects of both her action and inaction, and chooses inaction for a particular effect to be produced, is there a moral difference between tampered and untampered choice environments? With such strong evidence about behavioral influences, are Carolyn's acts and omissions morally distant enough to imply different moral conclusions? In this subsection, I claim that they are not. This gives rise to the so-called evidence-based view, which posits moral closeness between intended action and evidence-based predictability of untampered choice environments.

Some direction in these matters could be supplied from the seasoned philosophical debate on doing and allowing. It is standardly held that 'doing' or 'acting' is more difficult to justify. In the case of non-reflective influences, a salient distinction between these two categories would allow us to say that effects resulting from calculated design retain a layer of wrongness that is absent in untampered choice environments. But to establish the distinction and its moral relevance, we should be able to say why entirely predictable effects that are purposefully left unchanged by Carolyn do not undermine autonomy in the same way as designed choice environments.

The acts-omissions distinction has not yet been adequately imported into the discussions on nudging and non-reflective behavioral influences more broadly. Here, I only take the first glimpse at what this debate could offer to considerations of non-reflective influence. Most moral insights in our discussions on acts and omissions are drawn from thought experiments inspired by the famous trolley problem, which goes as follows:

> A trolley is racing down the track when its breaks become inoperative. Following a sharp curve, it comes across five workers working on the track, who are unable to avoid it. You are standing next to a lever that can divert the trolley onto another track. If you do, the trolley will run over a single worker, working on that track.[54]

The trolley problem explores whether there is a significant moral difference between doing and allowing harm. Over the years, the examples that are meant to test our intuitions regarding the difference between doing and allowing have taken on a variety of forms. To mention but a few, there are questions whether it is permissible to push a fat person off a bridge to, *ex hypothesi*, stop the train in the tracks in order to save the five, or whether it is permissible to harvest a healthy patient's organs in order to save five who are in dire need of a transplant. All the various examples capture the essence of the original one, but the debate compares them in search of distinguishing

[54] The example is adapted from Foot (1967) and Thomson (1985).

features that would explain different moral intuitions. For instance, more people would permit that the lever be pulled in the original example than they would permit healthy persons to be harvested for organs.

A well-known Kantian proposition for dealing with the differences between these two cases specifically is that unlike in the lever-pulling case, the transplant case sees a healthy patient being used *as a means* so that other patients can be helped. Judith Jarvis Thomson explains that for us to be used as a means in these cases entails that, if we were completely obliterated from existence, the five could not be saved. If there was no healthy patient, the doctor could not harvest organs and help the others. In the original case, however, obliterating the single worker from existence does not constitute saving the five (1985, 1401-1402).

Let's try to apply this to the cafeteria case. Can the 'using-as-a-means' condition explain the difference between actively changing choice environments or allowing them to have predictable effects? Surely, Carolyn could be said to use the heuristics of her customers, or themselves, as a means when she aims to change their behavior. But if this is correct, it is not obvious why allowing predictable effects of untampered cafeteria arrangements does *not* entail using the heuristics of persons as a means. If non-reflective influences are inevitable in such cases, and can be predicted by Carolyn in all available scenarios, then it would be inevitable for Carolyn to treat her customers as a means. Acts and omissions in such cases seem "morally close". Granted, this is only one distinguishing feature that we could apply to considerations of non-reflective influences, but any alternative would have to have enough traction to overcome the moral closeness.

This means that if a highly-skilled behavioral expert like Carolyn could, *ex hypothesi*, successfully predict the effects of all context-specific non-reflective influences on choice, then the

addition of intentional action, as opposed to "mere predictability" of the untampered choice environment, seems to carry little moral weight for autonomy considerations. Hence, if the closeness between acts and omissions cannot be overcome, we are faced with a new version of inevitability. And if this is the case, then Carolyn should look for non-reflective influences that are most compatible with her targets' autonomous pursuits. This is the core of the evidence-based view.

Another important consideration is that behavioral influences could be foreseen as a secondary consequence of primarily intended effects. Consider the following example by David Birks and Alena Buyx:

> "[…] imagine we paint our house green because it is our favourite colour. As it happens, the colour green has a soothing effect and reduces our neighbour's desire to be aggressive. It seems implausible that this harms our neighbour's mental integrity, despite the fact that it alters his desires. Whereas we might think that there is a moral difference if we were to paint our house green with the intention of changing our neighbour's desires." (Birks and Buyx 2018, 138)

Imagine a third case: when painting our house green, entirely due to our preference for this color, we also have evidence that our neighbor will become less aggressive. Or consider now applying a similar thought experiment to the cafeteria case: imagine that, instead of steering her customers towards healthy food, Carolyn is a benevolent employer whose primary intention is to arrange food so that the weight of food each cafeteria worker has to carry on a daily basis is minimized. Yet, being also the knowledgeable choice architect, it is predictable to Carolyn that this arrangement will, *ex hypothesi*, increase the likelihood that her customers will act on their strong cravings for unhealthy food, contrary to their goals. Does Carolyn wrong her customers by ignoring the foreseeable secondary effect? It seems that she does. This may be due to the fact that Carolyn is occupying a role in which, although not primarily intending a change in customer behavior, she bears responsibility for foreseeable secondary effects with predictable outcomes.

Let me then spell out two additional suggestion about the IA at this point: 1.) Designing the choice environment and leaving it untampered can be morally "close", when the choice architect can predict the non-reflective effects of both; in such cases, choice architects cannot avoid "factoring into" the autonomy-relevant influences on agents. 2.) When foreseeable non-reflective influences are a side effect of the choice architect's action, the choice architect can be held responsible for how they influence behavior.

My suggestions about the IA and autonomy have some limitations. First, they only seem to apply to cases where we expect each available choice environment to contain non-reflective influences of comparable brute properties. Examples of such categories of behavioral influence could be status quo biases, framing effects, or salience effects (as in the cafeteria nudge case). In other cases, these conditions might not hold. For instance, imagine that the aggression-inducing effects of the color red are much greater than the soothing effects of the color green. To respect autonomy would presumably involve attempting to remove the stronger kind of influence. Second, choice architects seem to bear responsibility for foreseeable secondary consequences primarily when they occupy particular roles. Of course, it is possible that others bear responsibility when autonomy could be seriously harmed by their behavioral influences. Imagine, for instance that Jane could paint her house in a soothing shade of green, but she knows that her next-door neighbor suffers from a condition that puts him into a lethargic state when exposed to such a color. With this knowledge in hand, it could be claimed that Jane acts impermissibly even if the non-reflective influence on the neighbor is a secondary effect.[55] Tying the responsibility for foreseeable secondary effects to expert roles merely reflects our expectation that expert choice architects like

[55] Some might suggest that Jane does not act impermissibly when she paints the house because she has property rights on her house, which include the option to paint it into lethargic green. But it would seem to me question-begging to impose property rights that are insensitive to new findings about how using our property may cause harms. Any theory of property rights sets the boundary for such rights in terms of when the use of property imposes harms upon others.

Carolyn will more often possess the knowledge that brings about the foreseeability of the secondary effect.

2.4.3. ***The advancement of behavioral science***

We might think that moral closeness between acts and omissions obtains only if the behavioral expertise about influences and their applications is so great that effects are reliably predictable to the choice architect. If, contrary to this, Carolyn could only predict the effect of her intervention, but not of the choice environment she replaces, then there could be an autonomy-relevant difference between acts and omissions. Some would probably object that to even conceive of such expertise is too "science-fictional", especially given Grüne-Yanoff's measurement concern. We could respond to the sci-fi objection in one of two ways. The first is to deny the unattainability of sci-fi expertise, by pointing to rapid advancements in the behavioral sciences, and to helpful insights about preferences we now draw from predictive and user-behavior analytics (so-called 'big data'). The dispute in this case would be strictly empirical. The other way is to claim that expert predictions only need to be sufficiently reliable for the IA to hold. This would mean to deny that the epistemic conditions of expertise need to be so high.

If either strategy is successful, then I believe that the advancements in behavioral science will strengthen the IA and favor permissibility of nudging. This is because, if behavioral advancements make it more likely for experts to accurately foresee the effects of non-reflective influences, then interventions on the choice environment (acts) will be morally close to allowing predictable effects (omissions). Since experts will have to do one or the other, a new kind of inevitability will thereby be introduced. What follows for the permissibility of influencing individuals in decision-making domains that they want to handle reflectively? It should follow that choice architects should at least try to substitute non-reflective influences that are difficult to resist

with others that are not. If even this is unavailable, then choice architects should be allowed to help individuals to achieve alignment between first-order and second-order preferences.

My opponents might propose that different normative lessons should be drawn from the advancement of behavioral science, which brings about the closeness between acts and omissions. Two proposals come to mind. If we believe that establishing closeness between acts and omissions with predictable effects is morally undesirable, then we might think that we should either 1.) suspend our work in behavioral science and purge as much behavioral evidence from collective memory as we can, or 2.) keep behavioral experts away from decision-making roles in which they could not avoid utilizing their expertise. Without delving too deeply into either proposal, I suspect both to be normatively problematic for reasons other than safeguarding autonomy, such as freedom of scientific research, scientific progress, censorship, and the dangers of keeping scientifically-versed individuals away from positions of power. Most of all, both proposals seem completely unfeasible in the modern world, which alone should be enough to cast them aside. This is not to say that we should not exercise caution when applying behavioral evidence, which is still far from sustaining the closeness of acts and omissions that I here described.

However, one stronger objection is available against designed, as opposed to untampered choice environments. Suppose, for a moment, that we have developed sci-fi expertise and that we can reliably predict the behavior of individuals when they are influenced non-reflectively. If you agree with my previous points, this would be enough to establish moral closeness between acts and omission and the new version of inevitability. Suppose that Carolyn, a champion among the sci-fi scientists, chooses between changing a cafeteria layout in one of many ways, or leaving it as it is, thus predictably steering behavior. But suppose also that Carolyn tinkers with a wide range of other choice environments. Although there seems little difference between Carolyn's acts and

omissions, there is one moral advantage to choosing the set of predominantly untampered influences, as opposed to the set of predominantly designed influences. In a less designed choice environment, influences will be more "scattered" in terms of their behavioral directions, instead of being more unified in direction.

How is this relevant to autonomy? There are two related ways. First, while pervasiveness means that agents will not be able to fully govern themselves, we might still care whether they are governed *by others* or *no one in particular*. Designed choice contexts with unified directions (towards, say, health promotion) seem to at least somewhat capture the sense of being governed *by others*, unlike untampered choice contexts. In this sense, designed choice contexts are more autonomy-threatening. Second, designed choice contexts with unified directions might make it more difficult for us to resist them if we disagree.[56] This observation could in fact slightly loosen the moral closeness between intended interferences and untampered environments with predictable effects. We could agree as citizens that intended interferences are more undesirable than untampered environments, if this produces choice contexts with more unified moral directions. But notice that this stands only *as long as* influences pull in unified directions. We should make sure individuals are not smothered by nudges pulling towards the same goal. This will contribute to the argument against stacking nudges, that I will defend as part of a system of nudge transparency in Chapter 3. It is to explaining this system and its normative assessment that I now turn.

[56] This is contingent. In Chapters 5 and 6, we will notice that untampered choice contexts in some areas are fairly unified in their heuristic pull. This explains why people do too little in areas such as environmental protection or poverty alleviation.

Chapter 3: Transparent Nudges[57]

In the previous chapter, I have introduced an account of personal autonomy that grounds our upcoming normative considerations of nudges. From this account, we can start drawing the picture of how various kinds of intentional non-reflective influences by government – nudges – may promote autonomous pursuits, but also how particular management styles might be threatened by them. A central concern with nudges is that as heuristics triggers, which are thought to interact with agents non-reflectively, they "typically work better in the dark" (Bovens 2009, 209). For instance, the cafeteria food arrangement may be found less effective if telegraphed to the intended targets. Opponents of nudges point out the concern that because nudges rely on psychological quirks, they "will be more effective if they are not transparent to the individuals subjected to them" (Grüne-Yanoff 2012, 637). Transparency is primarily a concern about the purported (il)legitimacy of government policies as it is intertwined with fundamental notions of accountability, respect, deliberation and consent (Hansen and Jespersen 2013, 15). Governments that nudge people 'behind their backs' without the latter being (able to become) aware of the influence are considered to be overstepping serious moral boundaries. Using such policy tools does not allow for the kind of scrutiny and contestation that we believe should be possible in liberal democracies. It is to the conditions of transparency that a nudge regime has to satisfy in order to be least threatening, and most enabling to personal autonomy, that I now turn.

In this chapter, I develop an account of nudge transparency, which enables analyzing and finding ways of meeting nudge critics' concerns, as well as exploring the institutional conditions in which nudges are compatible with personal autonomy. The aim of the chapter is both conceptual

[57] A large part of this chapter is adapted from my paper 'Nudging, Transparency, and Watchfulness' (2019) (Available at: https://www.pdcnet.org/soctheorpract/content/soctheorpract_2019_0045_0001_0043_0073). I thank *Social Theory and Practice* and my co-author Bart Engelen for allowing me to modify it for the purposes of the dissertation.

– to define and analyze what transparency in nudging entails – and normative – to assess how transparency of a nudge affects its permissibility within a regime of a behaviorally-enhanced liberal democracy. If convincing, the account results in nudging being advocated as a social pre-commitment strategy through the medium of government. First, I illustrate the predominantly autonomy-related worries critics have about the non-transparency of nudges and show why the standard responses fail. Second, I address the empirical question concerning how transparency relates to the effectiveness of nudges. Third, I build on insights of Luc Bovens and Andreas Schmidt to explicate the meaning and typology of nudge transparency in order to locate the normative concerns more clearly. Fourth, I reduce the scope of my inquiry into the permissibility of different kinds of nudges by setting uncontroversial cases aside. Fifth, I analyze the pivotal notion of *watchfulness* and develop a conception that is both feasible and acceptable in liberal democracies like ours. Sixth, I investigate the normative implications before rebutting some objections in the last section.

3.1. The worry, the responses, and why they do not convince

One of the main worries about the project of behavioral enhancement is that by playing into less-than-rational psychological mechanisms, governments influence people's behavior 'behind their backs' (Hausman and Welch 2010, 135; Bovens 2009, 216; Rebonato 2014; Waldron 2014). By doing so, governments covertly steer citizen behavior without this being obvious to the latter, which raises at least intuitive concerns about nudge permissibility. Importantly for my account, some critics argue that nudges are impermissible policy tools because they violate personal autonomy. It is the quality of being covert that could make nudges more problematic than

their coercive yet more forthright counterparts, such as mandates and sanctions (Ashcroft 2011). Similarly, Riccardo Rebonato argues that nudges have a greater accountability deficit, "exactly because the means employed […] are not transparent" (2014, 360).

In response to these worries, Thaler and Sunstein develop two counterarguments. First, they stress a familiar stipulation – that resisting nudges should be "easy and cheap" (2008, 6). If the individual perceives a nudge as irreconcilable with his long-standing goals and values, he can go against its influence. Why criticize nudges if they can be easily resisted? Opponents, however, are not convinced, since resistibility seems to depend at least in part on transparency. If non-transparent nudges steer choices "while flying below the radar screen of rational deliberations" (Rebonato 2014, 366), people may not realize they are being influenced and make decisions they would standardly reject. A lack of transparency makes it harder to sidestep a nudge, thereby potentially inhibiting people's reflective pursuits in light of their own management styles of pursuing a conception of the good. Nudges constrain reflective decision-making and thereby "prevent the reflective act of will required for a decision-maker to avoid the nudge consciously and willfully" (Mills 2018, 407).

Second, as I explained in the Introduction, Thaler and Sunstein are inspired by Rawls when they claim that governments wanting to employ nudges should do so publicly, with officials being "happy to reveal both their methods and their motives" (2008, 245). According to Sunstein, nudges "should be visible, scrutinized and monitored" (2015b, 148) in order to reduce the likelihood that they have illicit ends, reduce welfare or violate people's autonomy (2016b, 42). According to Chris Mills, nudges that are in line with people's ends, avoidable, made public and transparent, need not violate autonomy (2015, 502). Some critics agree that nudges can be legitimate as long as they are implemented transparently: "the more clearly visible, the better" (Rebonato 2014, 392). Others

remain unconvinced and argue that "the espousal of transparency and publicity constraints comes across as an artificial and ad hoc declaration of values that belies a lack of real interest in the importance of ensuring that those subjected to these subtle forms of state power understand the underlying rationale" (Anderson 2010, 374). Nudge techniques exploit cognitive and motivational heuristics – instead of explicitly mandating and sanctioning individuals – and thus seem extremely convenient to governments not wanting to face full scrutiny. Disclosing information about nudges, these critics argue, does not address the problem that nudges fail, by their very nature, to respect people as autonomous and rational persons (Waldron 2014). In other words, "we cannot be confident that publicity and transparency in combination would take away the government's motive and opportunity to manipulate" (Wilkinson 2013, 344). And even if we could ensure that governments nudge citizens with only the best of intentions, covert nudge techniques still seem objectionable for not living up to democratic standards of accountability, deliberation, and contestation.

There are two main reasons why critics remain unconvinced by Thaler and Sunstein's standard response that nudges are permissible when public. First, as I hinted in the Introduction and as the next section shows, critics like Grüne-Yanoff believe that, when disclosed, nudges will no longer be effective; while disclosure does not render them impermissible, it takes away their very purpose (2012). Second, Thaler and Sunstein's accounts of publicity and transparency are underdeveloped; they never offer a complete account of nudge transparency and what exactly it consists in (Bovens 2009). In this chapter, I address both issues. I show that most heuristics-triggering nudges are permissible and do not offend autonomy as long as they are transparent and can be resisted as a result. For this, I will need to provide a complete account of nudge transparency, thus addressing the second issue and rising above the shortcomings of Thaler and

Sunstein's account. Before turning to these tasks, however. I use the following section to assess the real impact of transparency on nudge effectiveness.

3.2. Transparency and the effectiveness of nudges

What exactly do we know about the effects of making nudges transparent? The most common expectation is that this impact will be negative. Bovens, for example, hypothesizes that the more specific nudgers are when disclosing information about nudges, "the less effective these techniques are" (ibid., 217). Grüne-Yanoff agrees that this will likely be the case, either because people will come to realize and resent that they are being steered and look to resist, or will come to see through the behavioral techniques, thus neutralizing their influence (2012, 637-638). If I come to understand how the scary pictures on cigarette packages work, "I will no longer find the drastic slogans and images shocking. Thus, the effectiveness of the policies requires their being not fully transparent" (ibid., 638). It is exactly the effective nudges that are rendered ineffective once transparent that I will attempt to incorporate here into a behaviorally-enhanced liberal democracy.

However, scarce empirical evidence suggests that transparency does not always decrease nudge effectiveness. A study by Loewenstein et al. shows that informing people about the use of defaults in steering decisions about advanced directives does not significantly weaken their impact (2015, 40). Transparency, then, does not increase resistibility against some covert influences. However, in response to Loewenstein et al., Sunstein rightly points out that individuals might react to the presence of defaults differently if the disclosure helped them to fully appreciate the impact

of defaults on behavior (2016c).[58] Furthermore, it is yet to be determined whether transparency changes the effectiveness of other kinds of nudges. That this will likely be the case is suggested by an experiment showing that attempts at exploiting reflexive reasoning can sometimes be overcome when individuals are merely instructed to carefully reflect on their decisions (Amir and Ariely 2007, 146-148).

Sunstein believes that the available empirical evidence still allows various outcomes to result from nudge transparency. Transparency might: 1.) reduce the effectiveness of nudges (because the induced reflectiveness allows individuals to identify underlying directions that they disagree with); 2.) make nudges counterproductive (because people show reactance if they do not like being nudged); 3.) make nudges even more effective (because people support the underlying goals) or; 4.) have no impact on effectiveness at all (2016b, 154).

One venue for further empirical research is whether effectiveness depends on the trust of the citizen in the nudger. When such trust obtains, one can intuitively expect the citizen to more easily agree with the direction of the nudge and view it as supportive of his goals. It is well-known that these aspects are crucial in generating public support (Sunstein 2017, 38), and it makes intuitive sense to claim that transparency would increase the perceived legitimacy and, thus, the effectiveness of the nudge. If you want to eat healthy and you are invited to a party at your friend's place, then you will welcome him influencing you by placing the potato chips in a distant corner and telling you about it.[59] Also, highlighting the effectiveness of a nudge can increase its

[58] Sunstein extensively discusses the findings of Loewenstein et al. and stresses that, in the studies, people might not be focusing on the information or caring about it. See also Sunstein (2016b, 154-57).
[59] Reisch and Sunstein (2016) show that people generally welcome nudges targeted at policy goals that they support and implemented by governments that they trust. This evidence in itself, however, does not say anything about the impact of transparency on effectiveness.

acceptability, but there is no evidence that this is the case when one is transparent about the underlying processes (nudges triggering heuristics) (Petrescu et al. 2016).

More problematic, however, are situations in which making nudges transparent helps individuals realize that they are being manipulated by someone whom they do not trust, and leading them in a direction they do not endorse. Since people perceive nudges in these cases as illegitimate, nudge transparency can lead to 'reactance' (Sunstein 2016b, 119). According to Robert Baldwin, disclosing nudges is "as likely to provoke protest as to reassure potentially targeted citizens" (2014, 854). One study provides evidence for such reactance among people who oppose the government's use of defaults; in fact, some of these individuals would pick the option that the nudge supports, but only in the absence of the nudge (Arad and Rubinstein 2018).

Transparency would, however, render many nudges ineffective in one sense. If people agree with or take up as reasonable the behavior that a transparent nudge helps them pursue, then the intervention no longer operates 'qua nudge' in the narrow heuristics-triggering sense. Its effectiveness no longer comes from the behavioral trick it employs, but from the information it conveys, which interacts with our reflective capacities.

As I will argue in this chapter, making nudges transparent meets potential concerns about autonomy on the resource-based account that I proposed in Chapter 2. On one hand, when disclosure renders nudges less effective, it probably means that individuals come to see the nudge as contrary to their reasons, goals, and/or management styles; this is the primary purpose of nudge transparency. On the other hand, when disclosure does not make nudges less effective, there are two possible explanations. Either the nudgees perceive the nudge(r)s as pursuing legitimate goals – which is morally unproblematic – or the particular nudge may not be easily resistible even when made transparent – which does raise moral worries. The main purpose of the chapter is to analyze

the conditions under which less-than-fully-transparent or non-transparent nudges, that only work qua heuristic triggers, can be justified. I argue for a technically elaborated account of transparency that does not require full disclosure of such nudges in the most standard sense. As I will show, there are normative reasons for nudges to remain undisclosed and work as heuristic triggers, so that, for many consenting agents, they may facilitate the pursuit of autonomous goals non-reflectively.

3.3. Transparency: a typology

Arguing that governments should be transparent about nudges and the reasons behind nudging leaves open what kind of transparency is expected (Sunstein 2016b, 73). Bovens's seminal paper on 'the ethics of nudge' sets up the conceptual landscape for discussing moral issues surrounding nudge transparency. His central distinction is between *type* and *token interference transparency*. *Type interference transparency* stands for governments informing citizens in general terms that certain kinds of behavioral techniques will be used to increase individual welfare or solve collective action problems. These governments are thus transparent about the types of interventions that they are going to implement. According to Bovens, however, this "is not enough" (2009, 216). Subliminal messaging, he argues, becomes no more permissible if it is openly admitted, and there seems to be little or no difference between type interference transparency of non-transparent nudges and disclosing the use of subliminal messages. Therefore, type interference transparency is not demanding enough. This conclusion is shared by Baldwin: "general strategies of disclosure will be seen as doing little to change the semi-covert nature" of nudges that target people's less-than-conscious cognitive processes (2014, 854).

Conversely, *token interference transparency* requires transparency about each particular nudge intervention. A number of problems arise here. First, the effects of nudges that owe their effectiveness to their covertness could be dulled, making token interference transparency a non-starter (Bovens 2009, 217). Second, even if feasible, it seems absurd to require each nudge to be signposted. Given the fact that some influences are inevitable in a brute sense, token interference transparency seems excessive: "beyond a certain level of effort, it will become absurd" (Cohen 2015, 40). Lastly, such transparency is not required for other kinds of policies, like laws or financial incentives. At any given moment, there is a plethora of regulations in the background that are not being disclosed every single time they may become relevant.[60] Token transparency, therefore, is too demanding.

What would suffice, claims Bovens, is an *in principle token interference transparency*, which entails that "a *watchful* person would be able to identify the intention of the choice architecture and she could blow the whistle if she judges that the government is overstepping its mandate" (2009, 217; emphasis added). Bovens stresses people's (not necessarily fully realized) ability to notice a particular nudge when subjected to it. Reformulating this, one can say that an *in principle* transparent (or 'detectable') nudge is not always *de facto* transparent (or 'detected'). Perhaps it works best when not detected, and thus with people who are not watchful, but it is not impossible to become aware of it and defy its influence. This is what distinguishes an in principle interference transparent nudge from subliminal messages, which are undetectable even for the watchful.[61]

[60] See also Schmidt (2017, 410).
[61] Hereinafter, I refer to Bovens's three categories as *type*, *token*, and *in principle* transparency.

A recent contribution to the transparency typology, following Bovens, is made by Schmidt, who proposes the principle of *reasonable token inference transparency*, according to which "a watchful person is someone who, with not unreasonable effort and understanding, would be able to detect a token of a nudge and would comprehend the intention behind it" (2017, 410). The important desideratum of this principle, which lies somewhere between token and type transparency, is that it does not impose unreasonable burdens on individuals. On Schmidt's view, most nudges can be suitably transparent in this sense, especially compared to other kinds of policies, which might be much more opaque and complex.

While Bovens and Schmidt provide a useful framework, they never explain what exactly their notions of 'watchfulness' entail and what the implications are for nudge transparency. In what follows in this chapter, I aim to analyze in much more detail what kinds of capacities are required for people to watchfully navigate the behavioral influences of nudges, following or rejecting them as they see fit. Although my preferred conception of watchfulness requires people to become savvier for noticing nudges than they currently are, I keep expectations within constraints of feasibility. If the conditions for watchfulness are satisfied, I argue, in principle transparency is an attainable ideal that allows for the use of at least some nudges that work in the dark, without personal autonomy being undermined. I also argue that in principle transparency resembles and, in some cases, boils down to type transparency whenever watchfulness is in place. When watchful, it is enough for agents to be informed about nudge use at a more general level.

Before going into the arguments, it is useful to distinguish between three 'axes' of transparency. First, Bovens's distinction can be said to refer to different *degrees* of transparency along a continuum that runs from only disclosing general information (type transparency) to disclosing specific information about specific interventions (token transparency). To show what

lies in between the extremes, consider other aspects of a nudge that can be disclosed: the timing and place of its implementation, the goal or intention, the behavioral technique and its impact, or the presumed causal mechanism that runs from the interference to the outcome.[62] The more details disclosed, the closer we get to the token side of the continuum. Take the implementation of food arrangements in public cafeterias. The government can be specific about one of the aspects (e.g., goal) while remaining vague about another (e.g., timing).

Second, transparency can depend on the intervention's *design*. Some interventions are fully *transparent by design*, while other owe their effectiveness to their covertness. But even influences that are transparent by design differ in terms of *when* the targeted individual notices them. In cases of *ex ante transparent* interventions,[63] the targeted individual sees them beforehand and can sidestep them if he so desires. Think of the urinal fly or traffic light labels (green, yellow, red) for healthy, less healthy and unhealthy food products. Conversely, an intervention is *ex post transparent* if the individual notices the influence only when it has already taken effect. Think of fake potholes painted on roads to slow down drivers, or the use of defaults in contracts (such as health insurance). Only after you come to experience the effect of the influence do you realize that you were nudged and that your status quo bias made an impact. Still other nudges are *non-transparent by design*. They are invisible to most nudgees, and only a well-trained eye can spot them. Think of framing effects, which are notoriously hard to detect, tapping into heuristics like loss aversion.

[62] Caution should be exercised about this last point. As I explained at length in the Introduction, Grüne-Yanoff (2016) argues that behavioral science does not adequately explain the causal mechanisms underlying behavioral change. Here, I do not go against Grüne-Yanoff. All I claim is that the findings in behavioral sciences should convince people that there are *some* mechanisms at play, many of which are in the lower capacity domain of cognitive processing.

[63] On a narrow conception of nudging that I am using in the dissertation, these interventions are not nudges proper. I come back to this shortly.

Third, transparency can vary for different *people*. The question here is whether nudges are detected or merely detectable depending on the capacities of the nudgee. A detectable nudge may actually be noticed only by some people. Think of governments trying to raise tax revenues by sending out messages that most people have already filed their taxes. Even though some may not notice the influence at all, others find it much easier to detect.[64]

Of course, these distinctions can be combined. A government can, for example, be type transparent about the nudges being used for a particular purpose (e.g., health promotion), occasionally token transparent about timing, placement and the techniques involved (e.g., the way food items are arranged), but not disclose information about the way the techniques are presumed to work. Or one can have a look at one technique (e.g., lines on the road) and notice that its timing and placement are *ex ante* transparent but its intention is *ex post* transparent, except for people who have heard about this nudge and its underlying rationale before.

3.4. Reducing the scope of inquiry

The above typology helps to clarify when transparency is most relevant and which cases can safely be bracketed from discussion as morally unproblematic. This section aims to reduce the scope of inquiry by putting aside techniques where transparency is not an issue for their permissibility. This allows focusing on the problematic cases in the next section.

First, let's consider interventions that are transparent by design. When I say 'transparent by design', I assume that the intervention is a means of interacting with an agent who possesses

[64] Not all nudges are like this. Even when disclosed, some nudges are notoriously hard to detect, and our inability to cope with their covertness gives good reason to avoid their use. More on this in 3.7.

minimal capacity for picking up on influences. As I show later on, the effectiveness and transparency of a nudge always depend, to some degree, on the nudgee's capacity. However, the threshold I start with here is very low. It is the capacity of an average Joe, who has no insight into the character of heuristics and biases, does not know how behavioral techniques are used, or to what extent his choices are susceptible to their effects. So, even if some agents have very low capacity, some intervention will be such that agents will be able, with minimal effort, to spot the intended shift in behavior and the means of achieving it (Hansen and Jespersen 2013, 20-21), *before* it has taken effect on behavior. Call these *ex ante* transparent interventions.

The famous urinal fly, for instance, has all the features of an *ex ante* transparent intervention. The visual stimulus provides sufficient content for individuals to consciously work out what behavior is being pursued (improving the aim of urinating) and by what means. Notice that on my definition of nudging, it is not evident that interventions like the urinal fly should be included. As I mentioned, the aims and means of the urinal fly are so obvious even to a non-expert that, if he invests minimal effort, the choice arrangement of the urinal fly virtually conveys reasons for action to him. If this is the case, and the agent can easily dodge the effect, then the urinal fly does not obviously fit into the category of non-reflective influences. Still, individuals often pay no attention at all to occasional influences like the urinal fly, and it is in cases like these that it quite obviously triggers their heuristics, i.e. operates qua nudge. If they are resistible, and if their numbers do not overwhelm agents, *ex ante* transparent interventions (that sometimes operate qua nudges) do not pose any real concerns for personal autonomy.[65] Given that agents can easily spot them, they can just as easily circumvent them, or interact with them reflectively, when they

[65] Both of these conditions will be discussed further later on.

represent weighty concerns. Such interventions may only work 'in the dark' or 'behind people's backs' when agents are not looking to invest any reflective resources.

Second, nudges can become transparent *ex post*, i.e., *after* they have taken effect. Take the use of shocking pictures of diseased lungs on cigarette packs (Baldwin 2014, 836). The effect is "more or less unavoidable to begin with, but transparent in a way that allows the influenced person to recognize the intention and means by which this is achieved as a direct consequence of the intervention" (Hansen and Jespersen 2013, 21). Baldwin describes such transparency in similar terms: "It is nevertheless the case […] that the target of the nudge would be capable, *on reflection*, of realizing that a nudge has been administered and assessing its broad effect" (Baldwin 2014, 836; emphasis in original). Here, individuals can notice being influenced after the fact. While they typically do not or may not realize that the influence is taking effect, they can notice it in the aftermath (*ex post*).

The potential impact on people's autonomy of *ex post* transparent nudges is significant. As I argued in Chapter 2, bypassing reasoning has the capacity to undermine people's personal autonomy. In this respect, *ex post* transparency cannot guarantee that individuals act autonomously, if this depends on their ability to circumvent the nudge, which is diminished if the influence is not transparent *ex ante*. If transparency is meant to ensure that nudges do not divert agents from pursuing their goals and values, as Sunstein argues (2016b, 42), we ought to either rule out *ex post* transparent nudges as impermissible, or try to turn *ex post* into *ex ante* transparency.

This last strategy might work as follows. Sometimes *ex post* transparent nudges become *ex ante* transparent to nudgees by mere exposure. A fake pothole may not have the same effect twice if the individual learns when and where to expect it, but he may fall for it if it appears elsewhere.

The first time you are influenced by a picture of diseased smoker lungs, you only realize it *ex post*. However, after being exposed to the nudge more often, you can see the influence coming and find it easier to resist. Empirical evidence shows how repeated exposure to a stimulus – like the kind on cigarette packs – diminishes the effectiveness over time, a phenomenon known in advertising as 'wear-out' (Strahan et al. 2002, 186). Other times *ex post* transparent influences become *ex ante* transparent when they have become culturally familiar to the extent that people rarely fall for their effect. Therefore, both anecdotal and more systematic evidence suggests that agents indeed possess a general capacity for converting *ex post* into *ex ante* transparency. And if some *ex post* transparent nudges could be systematically converted to *ex ante* transparency across society, they could be bracketed from further discussion.

Still, even with learning by exposure and by cultural habituation, a persisting worry remains that *ex post* transparency is insufficient for preserving autonomy. It is apparent that when individuals make decisions that carry particular weight to them and in which they want to minimize or overcome non-reflective influence, such as buying a house or making choices concerning education, nudges that are only transparent *ex post* seem problematic. It seems impermissible in these weighty cases that nudgees are exposed to an *ex post* transparent nudge, even if it happens only once. If we hope to turn *ex post* into *ex ante* transparency in such cases, society might have to pull extra effort to provide citizens with resources to learn where they can expect the instantiations of *ex post* transparent (and non-transparent) nudges. More on this later on.

Third, when stakes are extremely high, one can wonder whether transparency is required, since autonomy is not the most important concern in these cases. Take the nudging of drivers that aim at protecting those they endanger in traffic. Even in cases where these nudges infringe upon people's autonomy and may arguably constitute a *pro tanto* wrong, they can still be justified

overall because of the other values at stake.[66] Surely, making people comply with their enforceable duties and lowering casualty rates are amongst those values, as I elaborate at length in Chapter 5. Other vital concerns that may outweigh autonomy have to do with human rights, liberties and justice more generally. Consider nudges that would make it less likely for people to exhibit racist behavior or violate fundamental rights of others. Given the importance of such ends, transparency should not be required or regarded as a *conditio sine qua non* for democratic accountability and the protection of autonomy. I bracket nudges with such obvious and important advantages from this part of the discussion, since these overrule the autonomy concerns raised.

There are two slippery slope worries lurking here. The first is the worry of sliding into perfectionism about the kinds of actions that can and cannot be valuable expressions of one's autonomy. Without taking this worry lightly, I still insist that in at least some trade-offs between autonomy and other values, autonomy gets the short end of the stick. While some governmental measures – speeding laws and tickets, red lights, speed bumps, but also the smart design of roads – can certainly be said to reduce people's autonomy, they are justified for good reasons. These measures are perfectionist only if they are grounded in some specific form of human flourishing, which is only one of the many possible justifications – the most important one being the avoidance of serious harm to others.

The second worry is that, if autonomy takes second place, we might end up justifying more aggressive techniques for achieving desired social ends. Why stop at nudging? Why not use subliminal messages, hypnotize or neuroenhance people to honestly fill out their tax forms? Now, as Sarah Conly argues, there are good reasons to believe that stronger forms of influence *are*

[66] See also Conly (2012), Nys and Engelen (2017, 210-211), and Wilkinson (2013, 346).

permissible when other values clearly outweigh autonomy concerns, unless there are cases where it is particularly valuable that people autonomously act in deeply blameworthy ways (Conly 2012). If using these stronger forms of influence is not feasible or if non-transparent nudging is more effective in promoting the overriding value, we should resort to such nudging without a second thought.

I reach two normative conclusions about transparency in cases of conflicting values. First, if some value clearly trumps autonomy, then mechanisms that ensure *ex ante* transparency are not required. Imagine a scenario where fake potholes are an affordable and highly reliable way of slowing down drivers ahead of a dangerous curve. The lives at stake justify putting these in place without disclosure to drivers.[67] Second, if there is uncertainty about which value takes precedence, an in principle transparency constraint suffices and should be respected. In fact, nudges are particularly useful in this regard, since they allow agents to balance out values themselves, with transparency in place. Take an intervention that nudges people into charitable giving, such as a clearly communicated automatic enrollment of employers into donation schemes with an easy opt-out option (Cabinet Office Behavioural Insights Team 2013). If this strategy can be sidestepped when disclosed, then it is sensitive to the agent's pursuits. In a society committed both to alleviating poverty and respecting personal autonomy, people should be given leeway for balancing out the two values as they see fit at different times.[68]

Let me recapitulate. Since nudges are not a homogenous category, the purpose of this section has been to bracket cases where insisting on transparency is unnecessary: 1) when interventions are *ex ante* transparent, and resistible as a result; 2) when nudges that are transparent

[67] The matter of autonomy clearly being trumped by other values, and then the latter being facilitated via nudging, is elaborated in Chapter 5.
[68] The permissibility of nudging for charitable giving is the sole focus of Chapter 6.

ex post can be systematically converted to *ex ante* transparency, and can become resistible as a result; and 3) when other values obviously trump autonomy concerns. I will focus in what follows on *ex post transparent* nudges and *non-transparent* nudges that can be converted to *ex ante* transparency. These are nudges that critics are concerned with when they argue, for example, that "nudges can be far less transparent than command" (Baldwin 2014, 851).

What are the typical examples of nudges that are non-transparent by design? Baldwin mentions framing, which "can be used to shape the decisions and preferences of an individual in a manner that is resistant to unpacking" (ibid., 836). As mentioned in the Introduction, framing the risks of surgery in terms of mortality rates versus survival rates has a predictable influence on people's choices (McNeil et al. 1982). When subjected to such a nudge, agents will not typically come to realize that they are influenced, even after the fact. These nudges typically tap into shallow processes, making them hard to detect, unpack, and avoid.

Other examples are cafeteria food arrangements, decreasing the default size of food portions and drinks, or the use of social norms, the last of which were shown to be underappreciated by individuals as drivers of their behavior (Nolan et al. 2008). For most people who are not well trained to detect them, these nudges really do work in the dark. The questions concerning their permissibility and whether and how people can overcome them are central to my inquiry.

3.5. Watchfulness: importance and four conceptions

Perhaps the crucial consideration about transparency is whether transparent nudges are easily resistible. Saghai argues that a nudge is easily resistible if the nudgee "has the capacity to

become aware of" the influence and "to inhibit her triggered propensity" to do as influenced (2013, 489). Does this apply to *ex post* transparent and non-transparent nudges? According to Saghai, attention-bringing and inhibitory capacities can be activated even if influences are "'covert', that is, unannounced, and therefore not explicitly indicated to the influencee" (ibid.). How can this work? I define watchfulness in the following way:

> *Watchfulness*: the capacity of individuals enhanced by resources at their disposal to detect and resist *ex post* and non-transparent nudges.

Before I formulate a principle of watchfulness that allows citizens to navigate designed choice contexts and overcome other non-reflective threats to their autonomy in more detail, I lay out a number of reasons why watchfulness should be promoted throughout society.

Increasing watchfulness addresses the worry about nudges being one-size-fits-all policies in a world of diverse management styles and conceptions of the good life. In a society that lacks watchfulness, nudges indeed risk steering some people in choice contexts in which they want to invest their reflective resources, and in directions they disagree with. The potential of nudges to steer people towards their considered values and preferences is realized best if people can be watchful. Watchfulness enables citizens to avoid the nudges that are unlikely to facilitate reaching their goals or respect their management styles, while allowing them to set up or select their own choice architectures, which will steer their non-reflective behavior in ways they reflectively endorse.

While nudges have been said to reduce citizens to passive followers (Waldron 2014), watchfulness enhances their capacity to become 'planners' rather than mere 'doers' (Thaler and Sunstein 2008, 42). This both resembles and diverges from the idea of 'boosting'. According to

'boost' proponents (Grüne-Yanoff and Hertwig 2016), nudgers assume that people inevitably fall prey to heuristics and biases, and justify their tapping into these mechanisms to influence behavior. By contrast, boosts enrich and improve people's decision-making competences and skills. On my account, making nudgees watchful by teaching them about heuristics and biases does not assume passivity, but empowers individuals, quite like boosting does. But while watchfulness enables people to see nudges coming and circumvent them, boosts take it a step further, for example, by increasing peoples' statistical literacy and risk savviness (Gigerenzer 2014). Watchfulness is also less demanding than Peter John's 'nudge plus' policies, which encompass prompts to encourage slow thinking (John 2018). My approach has similarities to that of John, who also focuses on the need to democratize behavioral public policies through consultation and deliberation, but it is both more limited (I focus on more traditional nudges while John focuses on more demanding deliberative policies) and more clearly focused on transparency (which John largely takes for granted).

Instead of dichotomizing between empowering boosts and nudges that assume passivity, I instead suggest that nudges can be part and parcel of self-government in a liberal democracy. As Schmidt convincingly argues, nudges that are implemented transparently can democratize the control people have over their choice environments, thereby addressing the autonomy-related concern about nudges potentially increasing the extent to which people are subjected to alien control (Schmidt 2017). Watchfulness allows citizens to use nudges to their advantage when this benefits their life plans and management styles, when influences are inevitable and predictable to experts, or when this can counteract suffocating influences from the market setting. In a watchful society, doers (nudgees) are more likely to become planners (nudgers) and increase their control over their own behavior (via behaviorally enhanced environments). The watchful society

introduces a cooperative relationship between experts and laymen, since laymen can become aware of influences and decide for themselves which to resist, which to endorse, and which to create for themselves.

So, while the notion of 'watchfulness' seems promising in many respects, it is remarkably underdeveloped in the literature. I distinguish between four possible understandings of watchfulness. The first two stipulate fixed thresholds for the capacities that individuals are required to have in order to be watchful. The third isolates the high capacities to an expert few that are entrusted with alarming the citizenry. The fourth conception, which I defend, provides conditions that are conducive to enabling people to spot and circumvent nudges as they see fit. Each of these conceptions has implications for which kinds of nudges are permissible.

3.5.1. ***The minimalist conception***

First, according to a *minimalist* conception, the capacities of the nudgee – your 'average Joe' – are low so that ordinary people need not take special efforts to become watchful. While significant differences in terms of capacities can exist within the population, the focus of the minimalist conception is on establishing a minimal threshold that the vast majority of people supposedly cross. Nudges are then considered permissible as long as they can be unveiled by people of such a level of capacity. The attractiveness of this conception lies in its lack of demandingness. Taking people as they are, it places the burden of responsibility on choice architects for designing easily detectable nudges.

The unfortunate consequence of the minimalist position is that it rules out *ex post* transparent and non-transparent nudges, even when these generate highly desirable effects. After all, the average Joe cannot pick up on such nudges, even if people with some training in behavioral techniques can. Since the minimalist conception of transparency regards only *ex ante* transparent

nudges as permissible, it fails to cash in on the merits of nudges that work best in the dark. Given that *ex post* transparent and non-transparent nudges could play an important role in minimizing cognitive costs, as I explain later, the minimalist conception is too stringent.

3.5.2. ***The maximalist conception***

The second conception of watchfulness is a *maximalist* one: it sets the threshold for watchfulness at a much higher level. The basic idea is that people should have active awareness of the nudges that surround them. In order to keep the effective nudges that an average Joe fails to detect, this conception suggests that citizens develop a high capacity for spotting nudges that are not easily detectable. As a result, citizens detect nudges and adjust their behavior by either going along with them or rejecting them (Baldwin 2014, 840-842).[69] By contrast, low capacity individuals, like the average Joes, will "have very limited ability to adjust their behavior so as to reject messages that they disagree with and to act in ways that are inconsistent with such messages. They will, in turn, possess poor abilities to 'unearth' nudges such as defaults, and resist these" (ibid.).[70]

As mentioned before, moving from low to high capacity seems possible, as in cases where exposure helps turn *ex post* into *ex ante* transparency. Experiencing the same nudge repeatedly affects our capacities for picking up on its influence. However, citizens might need extensive training, which would require massive efforts from both their tutors and themselves. But, if feasible, this would relieve choice architects from being all too cautious as many nudges would become more easily spotted.

[69] Baldwin distinguishes between high and low capacity individuals, but does not commit to the maximalist conception.

[70] Baldwin's use of the term 'message' is inappropriate here, as cognitive techniques become messages in the proper sense *only if* they are unearthed. They are messages for those who can spot them, but not for those who fail to do so.

There are several problems with this view. First, it sets high demands on citizens, who are required to undergo extensive training and regularly update their knowledge about new behavioral techniques, regardless of whether this fits their management style. Second, it is questionable whether individuals can be educated to surmount certain kinds of non-transparent nudges. In particular, empirical studies have shown mixed results regarding our possibilities to overcome framing effects. Robin LeBoeuf and Eldar Shafir, for example, found that inducing reflection does not help: "framing effects are likely to persist even among careful thinkers" (2003, 89).

While the first two problems show that the maximalist conception violates feasibility constraints, other problems arise even if we assume people can in fact become maximally watchful. Why, one may wonder, do we still need nudges if such high levels of awareness and reflectiveness are attainable? This conception also disregards that citizens have good reason to welcome nudges exactly because they enable them to glide through some of their choices with little or no cognitive cost and thereby allow them to save up reflective resources for other activities. I would challenge the maximalist assumption on grounds that a permanent state of heightened attention is not desirable on quite a few management styles of autonomous individuals. Even if high capacities *can* be achieved, and individuals *could* effectively see through all non-transparent nudges, this is not what a resource-based autonomy demands or how many people want to lead their lives.

3.5.3. ***The whistleblower conception***

A third conception of watchfulness goes back to Bovens's notion of in principle transparency, according to which a *watchful* person could identify the intentions behind nudges and blow the whistle on government transgressions. According to the *whistleblower* conception, some people should be particularly watchful and alert non-watchful others (the average Joes). Here, the burden of watchfulness does not fall on everyone, at least not equally: the proper way of

fleshing out the government's duty of transparency is by effectively making it possible for a group of high-capacity individuals to publicly expose whenever the government engages in non-transparent and *ex post* transparent nudging.

This conception has its strengths, but it similarly fails to reap the benefits of nudges working 'in the dark.' As Bovens rightly stresses, it would be advantageous to have "some watchdogs with sophisticated equipment keeping an eye on the government" (2009, 217), but having whistleblowers is insufficient – not only some experts but all citizens should be able to detect nudges: "We find it important that also *we ourselves* could decide to become watchful and unmask any manipulation" (ibid.; emphasis in original). Following Bovens, I hold that a conception of watchfulness should do more to empower lay citizens in being able to raise the alarm on nudges that work below the radar. Whistleblowing for laymen may indeed be possible when non-transparent nudges work against particularly weighty preferences. Think of seeing a family member die because of a 'do-not-resuscitate' default rule for advanced cancer patients. Nevertheless, the majority of non-transparent nudges would remain detectable only to the expert few. The lay citizens' performance with such nudges depends on the successes or failures of experts to communicate their findings and to provide help for navigating past such nudges.

There are two other reasons why the whistleblower conception falls short of a complete principle of watchfulness. The first is the worry of nudge stacking: the danger of too many nudges suffocating the agent in his attempts to navigate influences (Coons and Weber 2013, 21). The whistleblower conception alone holds no constraints on how much nudging is to be permitted. Even if the whistleblowers end up competently revealing an endless sea of nudges, the nudgees would nevertheless find themselves burdened in their attempts to grasp all the ways in which their choice environments have been altered. Secondly, the conception places whistleblowers at odds

with the nudgers, quite like Wikileaks is a natural adversary to the US Government. While whistleblowers are indeed needed as a check on the potentially sinister ends of governments, we still miss out on the benefits of nudges in terms of their capacity to reduce cognitive costs. For this to be achieved, the relationship between whistleblowers and nudgers will have to go beyond an adversarial one and include a vibrant back-and-forth between all parties: government, experts, and lay citizens.[71]

To see the need to go beyond the whistleblower conception, recall the standard reflection-based worry that nudges as non-reflective influences could undermine people's control over their lives, or at least, on my account, in parts of their lives that they want to lead reflectively as best as they can. If transparency ought to help us in reflectively leading parts of our lives so as to best fit our conceptions of the good, then it is insufficient to only depend on the expert few to signal problematic choice environments. Citizens need to be able to exercise as much control as they can over nudges that they disagree with. But nudges are promising in a different way – they can enable personal autonomy by reducing costs on reflective resources in areas where citizens agree with the goals, and can thus focus on other choices that matter to them. As Sendhil Mullainathan and Eldar Shafir argue, people operate with a limited *cognitive bandwidth*. Faced with scarcity in time and resources, mired in pressing deadlines and worry, people often *tunnel* – use up their limited reflective resources and executive control for urgent tasks at hand. At such times, they find themselves more prone to error in other tasks and decisions (Mullainathan and Shafir 2013). As noted in the previous chapter, operating with a limited bandwidth for reflection may well mean that people will often be unable to effectively divide their attention to all the goals they wish to

[71] It could be argued that avoiding nudge stacking itself helps in overcoming this adversarial relationship. The more nudges there are, the easier it becomes for governments to hide illicit nudges from whistleblowers.

pursue, especially when they need to focus on pressing matters. Since non-transparent and *ex post* transparent nudges work below the radar, they do not incur costs by using up cognitive bandwidth. They can help in economizing attention and reflective resources, thus relieving people from having to divide their attention and aiding them in pursuing some of their goals effectively at low capacity. Nudges offer cognitive relief in another sense as well. Since nudging is a second-order strategy that delegates the design of choice architectures to qualified nudgers, "it exports decision-making burdens to someone else, in an effort to reduce the agent's burdens both before and at the time of making the ultimate decision" (Sunstein and Ullmann-Margalit 1999, 16).

3.5.4. ***The democratic conception***

I now turn to the fourth conception, which I defend, namely a democratic conception of watchfulness. It builds on the whistleblower conception, but incorporates opportunities not only for contestation, but for deliberation and participation. Nudgers can indeed cooperate with nudgees in the process of steering behavior towards ends that the latter would endorse. Here, experts act not only as whistleblowers, but also as consultants: they advise both government and citizens and help the latter to navigate through the nudges in place.

The idea is that citizens should have sufficient *opportunity for watchfulness* so that they can fluctuate between episodes of low and high capacity. This opportunity entails that citizens need not be conscious about and prepared for every nudge that influences their behavior. Instead, they should have resources readily available when some influence guides them towards actions they would want to avoid or that they would want to process reflectively, given their judgements about how they want to manage their lives. This means that they can choose whether to pick up on different types of nudges in different domains. Rather than noticing every single instance of

nudging, they should be able to notice the behavioral techniques the purposes of which they disagree with or those not fitting their individual management styles.

Watchfulness is thus a disposition, the actualization of which can be triggered at different times. First, a watchful agent with weighty preferences that go against some nudge can notice he is being steered in a direction he does not endorse. Saghai refers to evidence from cognitive psychology suggesting that "stimuli that […] produce a feeling of dysfluency are more likely to trigger scrutiny […] At least when individuals have strong and settled enough preferences, goals, or beliefs, they are likely to become aware of an anomaly". In such cases, people's "attention-bringing capacities" (Saghai 2013, 489) are activated and their disposition of watchfulness is actualized. Second, a watchful agent who is familiar with behavioral techniques can recognize them in similar or new choice settings, or he can look up the techniques on a governmental website when they affect an important decision-making area (thus converting *ex post* transparency or even non-transparency into *ex ante* transparency), enabling him to more easily bypass them. While the first remark is good news for nudge enthusiasts, the second aspect shows that there is some need to educate citizens.

The main advantages of this conception are twofold. First, it does not automatically rule out potentially desirable nudges working below the radar, and allows cashing in on their benefits. Second, it is realistic and not overly demanding in its expectations towards citizens. So far, I have argued that the democratic conception of watchfulness should be promoted because: 1) heuristics-triggering nudges are permissible only if there is such watchfulness; 2) such nudges are desirable because they promote desirable outcomes while preserving autonomy, by reducing cognitive costs and allowing people to pursue their conception of the good as they see fit. In the next section, I flesh out more fully what the implications are for a behaviorally enhanced, yet watchful society.

3.6. Watchfulness: normative demands

What are the precise normative demands that my conception of watchfulness poses? I argue that four conditions need to be fulfilled: an educational, a legal, a democratic, and a societal.

First, nudgees require a basic education for understanding and acknowledging the impact of heuristics, cognitive biases, choice environments, and behavioral techniques. Unlike the maximalist, this conception requires only a minimal understanding and acknowledgement of these factors as important drivers of everyday human behavior. This is also less ambitious from what proponents of 'boosting' try to achieve, namely a significant improvement of people's decision-making capacities. In addition, nudgees should be educated about the available information for looking up nudges that steer behavior in what they consider undesirable directions.

This is why the second condition stipulates a legal requirement, namely that governments publish a nudge registry: a resource readily available to citizens which compiles the nudge techniques implemented by governments (Lepenies and Małecka 2015, 435). For example, nudges and their goals could be published explicitly on a website such as nudge.gov.uk (similar to legislation.gov.uk – the online database of UK statute law) or in the Federal Register (the journal that publishes legal rules and notices of the US Federal Government). This addresses the worry about nudges being procedurally illegitimate if they are not "visibly brought into the legal system" (ibid., 433). It enables watchful citizens to understand which behavioral techniques their governments are using and why, and helps them circumvent those that are likely to conflict with their goals. As Robert Lepenies and Magdalena Małecka suggest, such a registry functions as a legal codification of nudges, increasing their visibility (ibid., 430).

Third, nudges need to be democratically legitimized. Citizens must be able to figure out what aims nudging governments pursue. Nudge strategies are often inherently tied to particular kinds of behavior, but not necessarily to particular aims. Consider the cafeteria food arrangement. Even if people do not agree with the government helping them lead healthier lives, they might agree that there is a collective duty to reduce health care costs. Knowing the aims pursued by nudges and having them justified in public helps people to focus exactly on those nudges related to behavior they find undesirable with respect to how they want to lead their lives.

Citizens usually acquire knowledge regarding policy aims not just through governmental disclosure, but through public debate. A public debate on nudge goals helps citizens become aware of nudges and coordinate their preferences with others in the face of expert information. Citizens can also change their opinions about certain influences in public debates, say, if they learn from experts that nudges in some area could substitute negative brute influences, or stacked intentional influences (such as those from the market setting, which I discuss in Chapter 4). On the side of the government, "citizen input can guide public decisions, which feeds into more responsive and efficient policies" (John 2018, 125). The exact institutional form of the communicative channels and deliberative fora in which nudge goals would be presented and discussed is uncertain, and will have to be determined in practice. The socially desirable form should succeed in gathering as many stakeholders as possible, tracking their long-term goals and raising weighty concerns. For instance, it has been suggested in the UK that behavioral influences are geared towards localized problem-solving (Burgess 2012, 5-6), which implies that debates on nudge goals should be organized locally. Still, given that people give most attention to political deliberation on the national level, it is not obvious that a localized debate reaches the greatest number of those affected.

The democratic and the legal requirement (a nudge registry) complement each other, if certain practical considerations are taken into account. I mention two here. The first once again points to the problem of stacking, i.e., to circumstances in which an overwhelming number of nudges paralyzes citizens in attempting to figure out how they work, why they are implemented, and where and when they can be expected. The democratic conception of watchfulness curbs nudge implementation to what can reasonably be expected of a citizen's attention span, and allows citizens to veto further nudge stacking through democratic debate and thus limit the number of nudges being implemented and listed in the registry. The second concern is that the registry should provide fairly stable and precise depictions of nudge techniques in practical contexts to enable citizens the successful identification of any intervention that they would want to resist. If similar techniques are used to steer behavior for different purposes, individuals may fail to successfully distinguish between them. In short, nudges in the registry have to be neatly illustrated and take fairly stable manifestations in public life.

Finally, the fourth condition, a societal one, incorporates the whistleblower conception into the account. In some cases, watchful citizens with superb knowledge of behavioral techniques or particular interest in a specific domain can and should warn others that unannounced non-transparent nudges are operational. Unauthorized influences might come to rise either inadvertently or as part of sinister scheming. Since a watchful expert can unfold such nudges and their underlying aims, and blow the whistle when necessary, a watchful society needs a sufficient number of such experts. They may be specialized in nudges within a particular field or with regard to particular forms of nudges.

Consider that physicians, in their interactions with patients, can use framing techniques (Blumenthal-Barby 2012, 361-363), which are particularly hard to resist for watchful citizens.

Imagine that physicians are required to do so as part of a governmental strategy to promote individual health. Even if nudge-educated and watchful citizens join public debates on health policy aims and there is a health nudge registry in place, they may have difficulties detecting and resisting such framing techniques. Here, the whistleblowers might provide a helping hand. Health experts can help keeping a watchful eye on nudges in health care, political experts can focus on agenda setting and formulating referenda questions, and so on.

As with other policy measures in well-functioning democracies, there need to be enough people with a strong interest in and knowledge of some particular issue to check, influence, and contest the activities of government. This provides a middle road solution between 1) a completely apathetic and uninformed society (where there is not enough actual watching going on) and 2) a completely active citizenry (where there is more than enough actual watching going on). In other words, it is a middle road between the minimalist conception of watchfulness, which is insufficient in light of the contestatory and participatory role citizens should have in democracies, and the maximalist conception, which is overly burdensome and unnecessary for well-functioning liberal democracies. In order to know whether a society is sufficiently vigilant, we need to look at the collective level, and check whether there are enough nudge experts and citizens active in different domains, and not just at the individual level, which is the focus of both the minimalist and maximalist conceptions.

Here I clarify the implications for nudges that are non-transparent (but in principle transparent) and *ex post* transparent to non-experts. Non-transparent and *ex post* transparent nudges are permissible only if 1) they result from democratic procedures, with citizens actively joining public debates, 2) the span of utilized nudge techniques is regulated (nudges are stable, not stacked, and appear in a registry), 3) experts act as whistleblowers and aid citizens in detecting and

navigating nudges when the latter see fit, and 4) they can be resisted as a result of 1)-3). If there are nudges that nudgees find detrimental to the pursuit of their goals or their management styles, they should be able to switch from low to high capacity and steer clear of their effects. As for areas where citizens agree with the intention, they can safely stay oblivious. In fact, citizens often have an interest in staying at low capacity in such cases, as that enables them to minimize reflective burdens in the face of a limited cognitive bandwidth. Humans rarely have interests in making all of their choices reflectively and unearthing nudges that steer their behavior towards ends with which they agree.

Bovens rejects type transparency since it is insufficient to make nudges permissible, quite like a general disclosure fails to justify the usage of subliminal messages. Instead, Bovens recommends in principle transparency, with watchful people being able to spot nudges. My principle of watchfulness adequately fleshes out Bovens's in principle transparency and why he rightly considers it to be valuable. However, I believe there might be little difference between in principle and type transparency in any watchful society. After all, Bovens's type transparency can be understood in different ways, with regard to what and how much is disclosed. What may be disclosed and democratically agreed on are any particular nudge's aims, techniques, and time or space frames. Consider, for example, a local government disclosing that it will start painting lines on roads on a particular date to minimize traffic collisions: this involves all of the components just mentioned. Type transparency may be compatible with remaining silent about any of these components, while still disclosing that some nudging will take effect and alarming nudgees to be on guard. But notice that, the more a nudge registry discloses, the more type transparency resembles Bovens's in principle transparency, provided the principle of watchfulness is incorporated.

3.7. Objections

Before moving on to the next chapter, I tackle a number of objections that might be raised against the account of watchfulness that I defended here.

The first worry might state that there is still something odd about conceiving autonomy in a way that its fuller realization is facilitated by nudges, which are essentially provided by the government as a walking stick for the attainment of our goals. This conception, one might argue, hardly captures the notion of people *governing themselves*. In fact, utilizing nudges in the way that I described could amount to governing ourselves *less* by being effectively steered by others. A true account of autonomy, the objection might proceed, would require reflective mastery over our choice environment, allowing us to act on reasons to further develop and revise our conception of the good. Delegating part of our choices to a nudge regime might be regarded as going in the opposite direction.

Although the objection might capture some of our standard intuitions about autonomy, it seems too demanding in the face of more recent psychological findings. Three responses are relevant here. Firstly, the coveted version of autonomy, one that ensures full mastery over the non-reflective elements, is simply irreconcilable with pervasiveness. Many of our choices, some of which strongly pertain to our conceptions of the good, are carried out at low capacity whether we like it or not. Reflection is a limited resource, and our remaining actions result largely from the workings of our cognitive heuristics. Oftentimes, we cannot avoid being influenced and steered even when we want to master a situation reflectively. This is the best kind of autonomy we can hope for. If the objection insists that it is not autonomy proper, then I can simply state that this is autonomy*, which still captures the notion better than other states of affairs. The revised account

of resource-based autonomy that I defend here allows individuals to balance out their reflective and non-reflective lives, and most, if not all management styles are compatible with some areas of decision-making being managed with the help of the choice architect. Secondly, watchfulness in a nudge regime, which has been my focus in this chapter, actually gives agents considerable reflective power over organizing their choice environments. Watchfully circumventing particular strands of influence helps agents to maintain considerable power over choice environments and to reflectively re-assess what is valuable about their conception of the good.[72] Thirdly, my account also allows us to think of nudges as social pre-commitments to certain kinds of behavior. With the help of watchfulness, citizens can use nudges to coordinate their behavior in ways they see fit, while respecting the autonomy of dissenters. In many cases, citizens will be able to agree that nudges preserve their autonomy very well in the face of strong brute properties of influences, the inevitability of choice architects, or the suffocating influences of the market setting. We can strengthen this point by drawing an analogy between individual and social pre-commitment. The analogy is reinforced because individuals can easily opt out of the nudge when transparency is in place. And since there is nothing wrong with people individually pre-committing to certain kinds of behavior in the pursuit of their autonomous goals, we should think the same for social pre-commitments.

Granted, the analogy here is moving a bit too quickly. Social pre-commitment runs into problems of monitoring, accountability and sinister ends, which do not arise with individual pre-commitment. My account aims to alleviate or even resolve such problems. Note, however, that the analogy also breaks down in ways that morally favor social pre-commitment. Remember that

[72] If it is argued that agents need to make a sufficient number of important choices reflectively, I need not disagree with this. In 2.3.2., I claimed myself that nudge projects should be modest and that there should be limits to how much individuals should be able to outsource their autonomy in order to retain self-government.

nudging, as a second-order strategy, can be more effective at reducing cognitive costs overall, given that the cognitive work of planning our decision-making at low capacity is carried out by other persons (Sunstein and Ullmann-Margalit 1999, 13). Therefore, it is not the case that if everyone is effectively carrying out individual pre-commitments, social pre-commitments can be eliminated at no cost.

The second objection raises the concern that making non-transparent nudges transparent by means of disclosure does not guarantee resistibility. Non-transparent nudges may be disclosed yet still remain irresistibly effective in steering behavior. Coons and Weber argue that disclosure in such cases may actually make things worse. If I were told that I would be coercively injected with a love potion, not only would the disclosure fail in making the act permissible, but could make it worse because I could notice my agency slipping away. Coons and Weber claim that some nudges could have this effect, like the (*ex post* transparent) distorting mirrors which make you look obese (2013, 20).

My response to this objection is straightforward. There is nothing in my conception that commits me to endorsing any nudge that cannot be resisted. Disclosing non-transparent and *ex post* transparent nudges only makes them permissible if they are thus made *ex ante* transparent, and thus resistible. With some techniques, like subliminal messaging (or, supposedly, love potion injections), this is not the case. Nudges are a diverse category, and some are certainly such that they can be made easily resistible with transparency in place. For any techniques that do not conform to these characteristics, governments have weighty reasons not to put them to use. Even

if some nudge is resistible to a vast majority, but nevertheless burdens a small minority that just cannot bring itself to resist it, gives us sufficient cause to suspend its use.[73]

The third objection asks why we should still rely on nudges if we can make people watchful. Why add governmental influences if there are already so many that a watchful society needs to unearth, ranging from pervasive biases to influences by marketers (I explore the latter in Chapter 4)? Sure, the objection goes, it is a good idea to increase watchfulness, but this should primarily help people to avoid the errors they already make due to manifold influences, not serve as an excuse to add government nudges and thereby increase the burdens of watchfulness.

Two responses are in order here. First, increased watchfulness about behavioral techniques and their impact can and should indeed help people more easily resist pervasive biases and influences by marketers that go against their goals and values. It strengthens their capacity to navigate through whatever influences they encounter (be they well-intended, less well-intended or not intended at all). Second, let me remind the reader that on my account, watchfulness is not the same as being completely reflective at any given time (a state of mind which would likely be undesirable on most management styles anyway). Given the limits of human psychology, a world with nudges will always have benefits that a world without nudges will lack. When choice environments are designed to successfully facilitate the choices people reflectively want to make, this frees up reflective resources that people can employ to focus on other choices.

The fourth objection stresses a distributive concern. Sure, democratically justified nudges may benefit the majority, whose choices are facilitated towards the goals they endorse, but they also impose cognitive costs on the minority, whose preferences and values go against the

[73] Nevertheless, this objection should give us pause. In many cases, the resistibility of a nudge, which depends on triggering particularly strong heuristics, or generating substantial social and emotional costs, will be difficult to assess.

implemented nudges.[74] Is it not unjust to disproportionately impose burdens: 1) on dissenters who oppose being nudged altogether, and 2) on people who do not endorse the directions of most nudges?

In response, it is not a feasible ideal for governments to try to design some neutral choice architecture that imposes equal cognitive burdens on all citizens. If any one-size-fits-all nudge policy disproportionately advantages some citizens at the cost of others (in terms of reflective burdens), then it makes sense for governments to focus on those nudges where they can safely assume what most people's goals are. With both high-stakes nudges (such as avoiding deaths in traffic) and low-stakes nudges (such as facilitating how to open doors by installing the right kinds of handles), these goals and values are largely beyond doubt. In both cases, nudges have strong legitimacy. In cases where a large part of the population disagrees with the direction of the nudge (Reisch and Sunstein 2016), nudges have low legitimacy, both in democratic terms, because they will lack consent, and in distributive terms, because they impose cognitive burdens on many. Remember also that if choice architects can predict that the effects of non-designed influences would create significant burdens for significant numbers, the choice of nudging may simply be the result of minimizing burdens across society. In such cases, the fact that some have to suffer cognitive burdens is justified because greater burdens would be suffered in the absence of the substituting influence. And even then, the burdened individuals are not left to struggle on their own, but are adequately redressed given all our investments into the conditions of watchfulness.

Let me take stock. In this chapter, I have laid out a principle of watchfulness that expands Bovens's analysis of nudge transparency. I argued that it respects personal autonomy, while

[74] This is a related but slightly different concern to the one voiced by Anderson (2010, 373), who stresses that "individuals vary in their capacities to resist […] This already raises concerns (nowhere addressed in *Nudge*) about the equality of effects that nudges have."

cashing in on the merits of various nudge techniques that work best 'in the dark'. I made specific recommendations about the educational, democratic, legal and societal conditions that need to be fulfilled for *ex post* transparent and non-transparent nudges to be permissible. I argued that with these feasible background conditions for watchfulness in place, in principle transparency advocated by Bovens comes close to type transparency. Also, this allows for a nuanced discussion of the permissibility of different nudge techniques while keeping an eye on democratic principles and constraints.

The normative suggestions that I provide here do not only sketch the contours of some utopian watchful society, but have implications for the legitimacy of nudges here and now. The conception is fully compatible with modest nudge programs, the implementation of which can progress with equally modest steps. This modesty of my proposal can be understood in two ways: in numbers (it avoids nudge stacking) and in force (it rules out irresistible nudges). My analysis thus strikes an adequate balance between the complacency of minimalism and the demandingness of maximalism. By increasing watchfulness, people are made more nudge-savvy, which helps them avoid impermissible nudges, without having to give up on the benefits of permissible nudges. Watchfulness empowers nudgees and thus addresses the worries of critics, while salvaging the benefits nudge enthusiasts aim for.

Chapter 4: Market Nudges

In the previous two chapters, I have introduced and defended an account of personal autonomy that is sensitive to behavioral findings, and then erected an institutional edifice, with transparency and watchfulness at its base, which protects personal autonomy and makes at least certain heuristics-triggering nudges permissible, while placing the domain of behavioral influence under democratic control. However, my argument for the permissibility of nudging from Chapter 3 has been constrained to cases where it is the government (or a coalition of government and private actors) that carries out the nudging.

Such an argument fits into the standard debates about nudge permissibility, in which governments act as the nudgers, and ordinary citizens as nudgees. This stems from the attempts of Thaler and Sunstein to promote nudges to governments as a new mode of policy-making. However, if the goal of the dissertation is to determine how and when our capacity to form, revise and pursue our conception of the good life is enabled or undermined by non-reflective influences, and if violations of autonomy are not limited to government interferences, then assessing the permissibility of government nudging is insufficient. The policy-maker's infrequent exploration of his new behavioral sandbox is a drop in the ocean of existing behavioral influences that we face in our attempts to pursue deep-seated values and goals. It is only appropriate, as part of a broader liberal assessment of behavioral influences, that we now turn our attention to the permissibility of behavioral techniques from a more common source – that of market agents.

Mainly for reasons of brevity, I shall refer to all behavioral influences used by market agents (hereinafter, 'marketers') in the range discussed so far in this dissertation as 'market nudges'. Some may find this imprecise or simply incorrect. It may be argued that once we extend the discussion to market interactions, we are leaving the nudge debate. The behavioral techniques

used in markets will not fit the conception of influences that are both mild and aim to steer people's behavior towards their benefit, by their own lights. This is a terminological matter. If critics are not satisfied with my broadening of the term 'nudge', they should interpret this as a similar strategy that I used for 'autonomy' in Chapter 2. They can assume that when I talk about 'market nudges', what I mean is 'nudges*', where 'nudges*' are deliberate behavioral influences that are not constrained by mildness or benefitting individuals by their own lights.

The chapter proceeds as follows. First, I comment on the absence of market nudges in debates on behavioral influence. Second, I adopt a simple morally significant categorization of market nudges and provide examples for them. Third, I claim that the use of behavioral techniques in a market setting raises significant autonomy concerns. I show that if critics are worried about autonomy being vitiated by nudges (as I explained it to be the case in 2.1.), their worry about market nudges should be greater, since 1.) market nudges are not constrained by principles of government nudging (mildness and sensitivity to agents' settled preferences), and 2.) are often "stacked" in great numbers that overwhelm agents. The verdict is that nudging in advertising, product selection, and market interactions in general is by far more morally objectionable than government nudging. Fourth, I respond to an objection that could be raised against my view, drawn from the work of Mark White, and discuss three further objections – pertaining to demandingness, conceivability, and site – that explain in part why market nudges have thus far been neglected in the literature. In the fifth section, I present several policy applications derived from my normative conclusions. My policy proposals consist of a three-part package: 1.) counteracting market nudges with government nudges; 2.) educating public officials to regulate the market setting; and 3.) introducing outright bans, information regulation, nudge-free zones, and other restrictive regulations.

4.1. Normatively engaging with market nudges

The presence of market nudge(r)s is acknowledged in *Nudge* very early by Thaler and Sunstein, when they raise the possibility that Carolyn, the behavioral authority, may simply utilize nudges to: "Maximize profits, period" (2008, 2). The book later sketches examples of behavioral biases by drawing on decision blunders in market transactions,[75] and argues for the permissibility of government nudges to counteract them (more on this in 4.5.). Thaler and Sunstein go on to claim that:

> "the invisible hand will, in some circumstances, lead those trying to maximize profits to maximize consumer welfare too. But when consumers are confused about the features of the products they are buying, it can be profit maximizing to exploit their confusion, especially in the short run but possibly in the long run too." (ibid., 240)

Still, in most places, Thaler and Sunstein's defense of government nudging seems to illustrate their purpose of counteracting biases generally, not counteracting the exploitation of these biases by third parties. The moral emphasis seems to be on the inability of individuals to get things right, and not on the tendency of marketers to exacerbate this inability.

Similarly, Christopher McCrudden and Jeff King criticize *Nudge* and later books by Sunstein for fixating on governments and consumers, while ignoring producers. They state that in "prominent examples, Sunstein largely sets to one side the *triangular* relationship of government, producer and consumer" (2016, 103; emphasis in original). One such example in the literature, that of New York's former mayor Michael Bloomberg restricting the size of soda containers for the hope of reducing obesity, is explained by Sunstein in terms of people tending to choose drinks

[75] Explaining the availability heuristic, Thaler and Sunstein state that in response to an earthquake, "purchases of new earthquake insurance policies rise sharply" (2008, 25). They do not explicitly say that government nudges should be advocated because marketers exploit the heuristic when selling the insurance policy, but this is heavily implied in later examples, such as that of credit brokers (ibid., 135).

in large containers and the government engaging in paternalism to counter their bias (2015b, 75-76), rather than "seeing the soda cup ban as restricting the harms caused by soda *manufacturers*" (McCrudden and King 2016, 103; emphasis in original).

As I claim in this chapter, it is morally erroneous to overlook the activities of 'manufacturers' in moral debates on behavioral influences. Several reasons ground this claim. First, marketers have a much greater cumulative capacity compared to governments to produce non-reflective behavioral influence. Second, they are much more experienced in influencing decision-making, and had been familiarized with many behavioral influences in practical terms long before these were researched and picked up by policy makers. Third, market nudges are not constrained by principles of benefitting those they target or helping to preserve autonomy, yet they influence people in areas where decisions are considered by most to be non-trivial. Richard Posner captures all three reasons when he states: "Businesses know, and economists are learning, that consumers are easily manipulated by sellers into making bad choices–choices they would never make if they knew better–in borrowing and investing, and in buying goods and services, such as food, health care, and education" (2013, 212).

4.2. Market nudges – understanding and distinctions

In this section, I analyze different ways in which market agents use behavioral influences. Importantly, the mere presence of behavioral influences will not be sufficient to raise moral alarms. As evident from Chapters 2 and 3, the presence of behavioral influences can be compatible with autonomous pursuits, even on accounts that are far more restrictive than mine. Furthermore, the analysis will be somewhat imprecise. This is due to the fact that the behavioral mapping of

influences across the vast range of market activities is still a new enterprise. It will take a while before we possess reliable ways of identifying influences and recognizing which heuristic becomes triggered.[76] Before the enterprise matures, we will have some normative difficulty with policy implications as well – should we crack down on market activities more generally (for instance, by banning advertising altogether), or regulate the market at a pace of trickling academic findings?

Possibly the most important distinction within the larger set of market nudges is between what I will dub *selling-a-nudge* and *nudging-to-sell*. Selling-a-nudge will entail presenting some behavioral aspect of the product as its selling point. Take, for example, Vitality's Glowcaps, the caps on bottles that glow in different colors when people need to take their medication. Similarly, since the explosion of Internet and smartphone technologies, the market has been flooded with apps that are designed to counter biases and help us stick to behaviors that we commit to, in a range of areas – diet, exercise, finance, procrastination, etc. Take HelloWallet, a finance aid that helps our commitment to improve savings by appealing to defaults, peer effects, and frames. This behavioral design is a selling point for products in that it is used to convince customers to buy this product instead of an alternative without that feature. Of course, there is nothing wrong as such with selling-a-nudge. Here, a nudge is sold as a commodity, and customers, having all the information at their disposal, consent to it when buying.

We enter more uncertain moral ground with *nudging-to-sell*. Nudging-to-sell denotes the marketer's non-transparent use of heuristic triggers to make it more likely that consumers will select a particular product or service. The understanding that marketers possess of behavioral

[76] This is further impeded by a resistance of sellers to disclose their methods. Wendel notes that "companies and marketing agencies that apply behavioral techniques for their clients and assess their impact are often under nondisclosure agreements about their work–they can't share lessons publically, and certainly not for academic publication" (2016, 98).

influencing may often be non-scientific and cultivated through market experience, but the techniques they utilize may in fact be more effective than those devised by behavioral scientists. The category of nudging-to-sell includes these techniques even if it is not specified how exactly they exploit cognition or whether they have been tested in controlled environments.[77]

A related, but perhaps not exactly identical category, marks the occurrence of consumers exhibiting some irrational behavior derived from low-capacity reasoning, but not under the influence of marketers. Still, recognizing their opportunity, marketers swoop in and profit from the irrationalities. Call this *preying on cognitive mistakes*. A podcast episode of Planet Money from 2014 (Episode 590: The Planet Money Workout) claims that a gym of Planet Fitness, one of the biggest gym chains in the US, only has the capacity for 300 people to exercise, but has 6,000 long-term subscribers. Half of these subscribers hardly ever go. Offering a monthly subscription is often sufficient for potential subscribers to make forecasting errors and commit to fitness plans that they cannot or will not see through.[78] The gym does not go out of its way to trigger the mistake in motivational forecasting,[79] but its subscriber numbers show that it is fully aware of the tendency and eager to cash in. Some might claim that cashing in on already triggered cognitive foibles is less of a violation by marketers than triggering those foibles. Still, the fact that cognitive errors, which may ultimately undermine autonomous pursuits, can represent incentives for profiteering

[77] My distinction between *selling-a-nudge* and *nudging-to-sell* could of course be further complicated by other grey-area categories. Two which come to mind are *nudging-to-sell-a-nudge* and *selling-a-nudge-to-sell*. *Nudging-to-sell-a-nudge* would entail the use of behavioral effects (anchoring, availability, priming, peer effects, etc.) in order to sell products that have potentially useful behavioral traits, which could have otherwise been used as selling points. *Selling-a-nudge-to-sell* would occur if, for instance, a firm sold its insights to other marketers on how to utilize behavioral techniques. The first category is an extension of *nudging-to-sell* and is thus morally problematic for the same reasons. The second should be considered *enabling* or *collaborating* the act of heuristic exploitation.

[78] For a detailed account of consumer behavior related to picking gym subscriptions and acting on them, see DellaVigna and Malmendier (2006).

[79] This is not to say that gyms, or any businesses that sell subscriptions do not engage in nudging-to-sell. For a list of methods, see Beggs (2016, 139-143).

should be reason enough to regulate the market in practicable ways, and bolster the recognition of consumers when their mistakes are exploited.

Nudging-to-sell and preying on cognitive mistakes can be expected in three domains of market interactions. The first is in our direct dealings with salespersons. From telemarketing to real estate, salespersons have been trained to steer verbal exchanges in less-than-rational ways in order to make sales. The methods used, as in all other domains that I will discuss here, can be divided into informational and psychological misdirection.[80] In car dealership, for instance, salespersons are mainly focused on making some information more, and some less salient: keeping the customer's attention on additional components (which are hardly ever used later on), zooming in on the price of the monthly payment while distracting from the length of payment, masking inspection and gas prices, etc. (Akerlof and Shiller 2015, 62-63). Other techniques are easier to pin onto the heuristics map. Take the practice of haggling, in which salespersons rely on anchoring. The initial price, often set high by the salesperson, prompts you in offering your own prices, and determines how you assess the appropriateness of working your way down.

Second, consumers are nudged in terms of how selection between different products is framed. As Dan Ariely points out, items on the market are picked, without exception, in comparative terms. This is because we "focus on the relative advantage of one thing over another, and estimate value accordingly" (Ariely 2008, 2). Still, the comparative character of reasoning about prices can be steered to favor some options at the expense of others. Ariely's well-known example about economist.com describes the decision between three subscription options – a web-only subscription charged $59, a print-only subscription charged $125, and a print/web

[80] This is adapted from Akerlof and Shiller's distinction between information and psychological *phools* (2015, xi).

subscription also charged $125. As anticipated, hardly anyone picks the print-only subscription, given that the print/web option includes the web subscription at no additional charge. Obviously, the print-only subscription is not on the menu to actually get picked, but to present the print/web option as 'a bargain' (ibid., 2-3). This is confirmed by Ariely's study, which tested, on 100 students, two menus of options, one including the print-only option and one excluding it. With the print-only option included, only 16 students picked the web-only option, while the remaining 84 went for the print/web subscription. Without the print-only 'decoy', 68 people picked the web-only option, while only 32 picked the print/web deal (ibid., 5-6).

Techniques using decoys are ubiquitous in the market setting. As mentioned in the Introduction, restaurants, for instance, use a decoy in the form of a particularly expensive meal, so that other slightly less expensive (but still profitable) dishes will get picked more often. Another inevitable feature when making market decisions is the physical placement of goods that we choose from. These are the cases where the supermarket Carolyns figure out ways to 'maximize profit, period'. George Akerlof and Robert Shiller note the strategies of supermarkets to place the milk and eggs in the back so that you have to traverse the entire store in order to get them (and presumably, buy more items as you go), and temptation items like candy, magazines and cigarettes at the counter (2015, 21).[81] Additionally, there are quite a few resources spent nowadays on determining how exactly shelf placement best captures the consumer's eye.[82]

And then there's advertising. Ads efficiently utilize the heuristics list, including positive and negative frames, statistical frames, the zero-risk bias, priming, the bandwagon effect, and

[81] Another common strategy in European supermarkets is the placement of fruits and vegetables at entry points, due to their significantly shorter expiration date compared to other products.
[82] For eye-movement studies that explore consumer behavior at the point of purchase, see Chandon et al. (2009), Henderson and Hollingworth (1999), Wedel and Pieters (2008).

anchoring. Often, ads create associations that are irrelevant to the product, say, with celebrities or sex. Or they point out an aspect of the product that can be expected in any of the competing products, thus playing on consumer ignorance. Moreover, Akerlof and Shiller note that advertisers have not only been honing their skills at utilizing non-reflective techniques, but have developed sophisticated methods to measure their success (ibid., 45).

One final domain that I will mention in this section, which in fact runs across the already mentioned categories but deserves a bit of separate attention, is pricing. I adopt the examples here from the list of so-called 'obfuscation' tactics that Robert Sugden has recently described. *Drip pricing* draws the consumer with low prices to the point of purchase, but the process of buying is revealed to have many hidden fees. *Baiting* is the practice of offering low starting prices that are maintained only while 'stocks last'. *Price partitioning* sells items by components, and thereby obscures the total price. *Buy-now discounts* set up time constraints for consumers to look for alternatives (2018, 157).[83]

We may wonder whether markets are conceivable without these elements. It seems impossible to avoid dealing with salespersons, or expect them not to be biased towards selling, just as it seems impossible to eliminate comparisons between two products or their prices. Furthermore, it could be conceded that advertising does have the capacity to undermine autonomy, but might only be *pro tanto* wrong, given its positive social function to inform consumers and thus increase their choice range.[84] These are all noteworthy points. The success of any normative argument I make here about the capacity of the market to undermine autonomy would have indeterminate

[83] See also Shell (2009).
[84] If this were the case, however, the burden of proof would be on the marketers to show that advertising carries greater benefit than the harm it produces.

policy implications. Our policy decisions may often favor current market practices, if alternatives only seem worse or more uncertain.[85]

Still, I can respond to this worry by making two points. First, some policy ideas are available to tackle at least some heuristic exploitation. We might think that the way supermarkets or option menus are arranged should not be completely at the discretion of the seller. We might also think that the ways in which we organize public spaces determines, at the very least, the extent to which the population is exposed to advertising; ousting advertising from some public spaces, for instance, would seem an available policy option. Second, it would be normatively sufficient to offer a strong autonomy-based argument against market nudges, and claim that it should guide future policy engineering. The argument would not be undermined just because there is no readily available policy formula. The two points do not have the capacity to protect consumers from autonomy-undermining market influences entirely, but will shorten the range of threats, which consumers can then try to overcome reflectively. I will tackle the conceivability problem at length later in 4.4.2.2., and will now turn to detailing my main argument pertaining to autonomy.

[85] Responding to claims that the exploitation of cognitive biases by credit card companies contributes to people running up excessive debt or declaring bankruptcy (see Bar-Gill 2004, and Sunstein 2006), Richard Epstein argues that banks and credit card companies actually have "a powerful market constraint against excessive borrowing" (2006, 127), and that the occurrence of bankruptcies are a regular occurrence if borrowing is allowed. Individuals could, in fact, demonstrate hyperbolic discounting or excessive optimism when they borrow from banks and credit card companies, but that itself, Epstein says, is not sufficient to argue against such borrowing. With no access to it, people would not be "denied all access to credit," like borrowing "at high rates from loan sharks" (ibid., 128). I do not mean to take a stand on the credit card borrowing debate here, but merely show that paving the policy road may often entail preserving current practices with their current (lack of) regulation, given the undesirability of alternatives.

4.3. Detailing the autonomy argument against market nudges

At the heart of the matter is whether market nudges undermine personal autonomy. I have already hinted that the rejection of market nudges can be derived from internalist objections that many have raised against government nudges. Recall these objections from Chapter 2. They suggest that nudges bypass reasoning, and thus effectively diminish the control individuals have over their deliberations, evaluations, and actions. Market nudges seem to trigger the same kinds of non-reflective cognitive processes. If the autonomy-related criticism against government nudges amounts to these counts – namely, reflection bypassing and control reduction – critics of nudging would have to conclude that market nudges are *just as worrying as* government nudges. Expectedly, autonomy-related concerns about market nudges run deeper than the standard internalist worry. I claim in this section that after all the relevant autonomy-related features are examined, market nudges turn out to be *much more worrying* than government nudges.

It should be noted at the outset that my deviation from the standard reflection-based account makes my account less susceptible to the internalist objection. The resource-based account of autonomy, which I have outlined and defended in Chapter 2, and according to which my case for permissible nudging in Chapter 3 is tailored, sometimes allows individual preferences to be non-reflectively influenced in decision-making areas where individuals do not wish to invest resources of deeper reflection. For many, the market, or its greater part, is exactly the setting where they wish to invest little cognitive effort.[86] Still, the complexities of people's deep commitments should not be assumed. People likely care a great deal that they are good rational reasoners in some

[86] Individuals could also be completely indifferent in certain cases of decision-making, in the sense of having no decisive preference. I take it that in most situations of this kind, individuals would not expend reflective resources, and that the cultivation of preferences by market nudging for these individuals would not be autonomy-undermining. I thank a participant of the Mancept 2018 conference for raising this issue.

of the more crucial market decisions,[87] or in areas that relate more closely to their conceptions of the good. Because people's conceptions vary significantly, I have claimed that government nudges do have the capacity to undermine autonomy directly (when they intrude on processes where individuals want to maintain reflective control) and indirectly (when they steer individuals away from desirable outcomes and force them to invest reflective resources where they would rather not). Market nudges seem no different.

My endorsement of resource-based autonomy also explains why I resist a push in the opposite direction. Some might think that the invasion of market nudges justifies the use of government nudges, or at least downplays autonomy-related objections against them. Indeed, it is hard to shake off the latter intuition – if there are so many market nudges around, then what difference do a few government nudges make; in fact, it could mean that nudges in general are not *deeply* or *gravely* autonomy-undermining. But the intuition about the decreased moral gravity of both government and market nudges might simply be based on a status quo bias against increased regulation,[88] or an ill-founded optimism that we are not very susceptible to market tricks and advertising. The intuition might show a lack of appreciation for just how impactful behavioral influences can be. The resource-based account of autonomy, I contend, manages to appreciate the impact of behavioral influences much better than such an intuition.

[87] Because some market decisions (profession, health, finance, living accommodation, vehicle) often carry significant weight for people's livelihoods, and may deeply affect their conceptions of the good, there is good reason to give most regulatory attention to these domains of market decision-making. I do not explore the idea further here, but it can be assumed to complement my policy recommendations in 4.5.

[88] Whether an attitude against increased regulation is founded on the status quo bias can be checked with the use of the reversal test (Bostrom and Ord 2006). Imagine we believed that regulation against market and government influences, which would decrease their number, has bad overall consequences. We could ask, in reverse, whether regulation of market and government influences that would, on net, increase their number also has bad overall consequences. If we answer in the affirmative, then it is likely our judgments are derived from the status quo bias.

I suggested that market nudges are much more worrying on autonomy grounds than government nudges. So, apart from the internalist objection, with which government and market nudges fare equally, there are further autonomy-related counts on which market nudges fare worse than government nudges. I mention two such counts here.

First, in the absence of coercive regulation, market nudging is not constrained by the principles of government nudging, on either Thaler and Sunstein's account, or my own. Accounts of justified government nudging are standardly constrained in general terms by avoidance-enabling mildness and a sensitivity to the nudgees' views about their own benefit.[89] Marketers will not look to provide individuals with opt-out strategies, nor will they subscribe to any rule that prohibits exploiting their weaknesses and steering them away from their considered judgments and deep commitments.[90] Whether market nudges will respect personal autonomies in these two respects – namely, the promotion of resistibility/avoidability and the deep commitments of persons – will depend on whether the economic incentive underpinning the nudge happens to coincide with such respects. When governments nudge, on the other hand, they must approximate the good of individuals to the best of their abilities, and devise easy exit options for dissenters.[91]

[89] I will work with the assumption that governments can reliably advance such goals, with watchfulness in place (see Chapter 3).

[90] Of course, this mostly relates to commitments that are at odds with some first-order preferences that have a propensity to be triggered in individuals. When there is no such propensity, individuals with a strong commitment will have an easier time resisting the influence. Recall Saghai's point from 3.5.4. that influences produce a feeling of dysfluency when they are at odds with strong and settled preferences.

[91] My claim that governments are more easily justified in nudging because they engage in such approximation is a serious point of contention for White. I extensively address this point in 4.4.1.
Additionally, the design and implementation of market nudges tends to differ from that of government nudges. Abdukadirov has made a strictly effectiveness-based claim that markets are a more suitable "playground" for coming up with and updating nudge techniques, given the greater feedback from targeted individuals, while governments are far less successful "because their proponents attempt to use a legislative process to develop consumer products" (2016, 182). However, if the interactive character of market nudge application confirms their greater cumulative capacity to steer action, then Abdukadirov's effectiveness argument may in fact run counter to an autonomy-based argument. Compared to governments, marketers will upgrade their know-how for steering behavior at a faster pace. This will increase their capacity to undermine autonomy, as well as their technical abilities to update their influences quickly enough to evade the counteracting government nudges (see 4.5.).

Market nudges do seem more tolerable when they jolt people to act on their deep commitments. Many market nudges are designed exactly with such purposes in mind. Following Jodi Beggs, I refer to market nudges that are both profit-maximizing for sellers and promote (at least most) individuals' settled preferences as 'Pareto nudges'. Conversely, by 'rent-seeking nudges', Beggs refers to techniques that maximize the profit of sellers, but run counter to the settled preferences of individuals (2016, 127).[92] Although Pareto nudges fail to appease internalists, and are not constrained by mildness in principle, they seem at least more respectful of personal autonomy than rent-seeking nudges. In fact, if the market provided sufficient incentives for marketers to carefully gauge the settled preferences of customers and stick to Pareto nudging, the autonomy problem would certainly be more bearable. Companies do often have strong prudential reasons to engage in Pareto nudging, but once again, these motivations are circumstantial. Pareto nudges will only be a sustainable method for sellers when prompting action that fosters deep commitments makes sense within a broader profit strategy. It will often be more profitable for marketers to tempt weak wills instead, and in so doing, undermine the synchrony between first-order preferences and deep commitments.

Some market enthusiasts will no doubt suggest that a healthy competitive market will punish rent-seekers. If consumers' cognitive weaknesses are exploited by marketers, they would claim, consumer demand will go to other, more trustworthy suppliers. In some cases, claims about the capacities of markets to self-correct might be plausible, and it will then pay dividends for businesses to build up a reputation of honesty. But the expectation that consumers will flock to honest businesses is naïve about the consumer experience. For one, consumers will often fail to

[92] I take both concepts to be intention-based, meaning that they are designed intentionally to promote either long-term interests, or exploit short-term temptations. Many designs of this kind will fail. Some Pareto nudges will fail to promote settled preferences, and some rent-seeking nudges will not be rent-seeking as envisioned. But in that case, they should be treated as *failed* Pareto nudges and *failed* rent-seeking nudges.

notice that they have been non-reflectively influenced, and even if they do, they may not recognize the influence as decisive in producing their action. Both tasks are undermined by an undue optimism of many that they are immune to market influences such as advertising, or at least significantly more resistant than others. Admitting to ourselves and to others that we have been successfully duped can be difficult and humiliating. It does not come easy to successfully monitor influences, or to blame nudgers for bad outcomes, and not ourselves.

The second, and possibly graver count on which market nudges fare far worse than government nudges is nudge stacking. In Chapter 3, I leaned on Coons and Webber's argument that excessive government nudging can suffocate individuals in their attempts to navigate the behavioral landscape. This gives us reason, I suggested, to keep the government's nudge projects curtailed in terms of nudge numbers. Otherwise, watchfulness, the purpose of which is to enable dissenters to spot nudges they disagree with, may well lose its effectiveness. Market nudges, which stretch along all domains of market interactions, are bound to be excessive in number. The various mechanisms pertaining to transparency, even if they successfully covered the whole range of market nudging, would not be able to do the work that they would do for tracking government nudges. The market setting produces strong incentives for agents to nudge, if they want to keep up with their competition. In fact, it seems that the healthier the market competition, the more market nudges are stacked, and the more difficult it becomes for consumers to navigate the behavioral landscape. Moreover, our modern environment is characterized by the advancement of new and expanding communicative channels, which provide new opportunities for market nudging. If circumstances are such that marketers enjoy an ever-larger platform for non-reflective influencing, and are incentivized to do so to keep up with their competitors, then we might believe market nudges pose the threat of additive harm. In other words, even if you believed that any single market

nudge does not have the capacity to threaten autonomy, you could still hold that the threat comes with their ever-greater numbers.

Interesting empirical questions could be raised by neuroscientists and behavioral scientists about how the brain processes the environment stacked with market nudges: Can our attention span successfully track influences over a multitude of market interactions? Do we suppress some of the advertising into the background and succeed in not noticing it? Or do we become overwhelmed, so that greater influence is derived from greater numbers? These questions caution us against making hasty judgments about how exactly nudge stacking relates to personal autonomy.

Instead, I here want to make three careful, but commonsensical suggestions. First, nudge stacking in markets seems relevant for autonomy because it increases the likelihood that individuals will be steered in directions that they disagree with in general. In other words, the more nudges there are in the market, the more likely their effects will not be resisted. Second, not only does market nudging entail more temptation in areas where individuals often feel weak, but the attempts at resisting these nudges, through learning where to anticipate them and being on the lookout for them, will include cognitive effort that individuals would often rather expend on other purposes. Third, a relevant point that I have ignored until now is the capacity of markets to tailor social values and norms. If there is a positive correlation between nudge numbers and the prevalence of behavior they promote, then market nudges will play no small part in such consolidation. Values and norms will often emanate from behaviorally-tailored market preferences, and later represent social pressures for individuals when they form, revise, and pursue their conceptions of the good. It is, of course, permissible for people to endorse such social norms and allow them to guide their preference formation, but as mentioned in the Introduction and

Chapter 2, social norms are also a non-reflective driving force of behavior, which individuals often fail to acknowledge.[93]

Some claim that it is in fact markets that enable the flourishing of personal autonomy. Most recently, John Tomasi has argued that self-authorship is conditioned upon individuals enjoying a bundle of 'thick' economic rights, which include the right to run one's business as one sees fit and the right to control one's assets without coercive interference (2012, 22-23). But assuming that these rights are conceived to allow unhinged market nudging, arguments like Tomasi's need to show why autonomy is not compromised by it. As Matthew Clayton and David Stevens point out, whether unrestricted markets will promote autonomy depends on whether the effects of people using their thick economic rights limits or expands the self-authorship capacities of others (2015, 365). In unrestricted markets, marketers have a strong incentive to utilize market nudges for monetary gain, thus undermining citizens' capacities for self-government. Additionally, in an unrestricted market, sellers can focus on designing their nudges to make them as effective as possible, while buyers are pressed to divide their attention between various purchasing decisions. If such a set-up systematically favors sellers over buyers, then market nudging is objectionable not only on autonomy grounds, but on grounds such as equality and exploitation.

Tomasi could complain that removing interferences for the buyers could diminish their sense of self-respect, if this means that "one is not (and so cannot think of oneself as) the central *cause* of the life one is leading" (2012, 83; emphasis in original). But it would be wrong to suggest that removing interferences in the form of market nudges undermines the status of persons as 'causes' of their lives. Quite the contrary – removing nudges eliminates other causes that could

[93] For interesting takes on markets shaping preferences, see George (2001) and Lindstrom (2008; 2011).

prove decisive in steering their actions, and that often run counter to their conceptions of the good. And even if it could be shown that people require market interferences to preserve self-respect, this might bear little moral weight compared to people's autonomies and other normative considerations.

4.4. Objections

I have claimed that market nudges have the capacity to undermine personal autonomy on the resource-based account I have introduced in Chapter 2. I have noted that the threat that markets pose to autonomy is more considerable than that of government nudging. This claim is not without opposition. First, I address the claims of White, whose argument runs in reverse to mine. But even if my argument is more convincing than White's, and we could establish the greater normative threat of market nudging, we may still be lacking clear normative implications for both marketers and public institutions. I will claim that there are two main questions that are thus raised: 1.) Are marketers acting impermissibly when they utilize profit-driven behavioral influences? 2.) Should we regulate market interactions so that behavioral influences are significantly reduced, or their effects neutralized? I dub these the *ethical question*, and the *political question*, respectively. Answers to these questions may pose ethical and political requirements upon marketers and public institutions. Attending to the further objections – pertaining to demandingness, conceivability, and site – will help shed more light on these matters, and prepare us for a short survey of policy suggestions in 4.5. Furthermore, it is these objections that made nudge theorists reluctant to engage

in either of these two questions.[94] I spend the rest of this section explaining why, and claim that some of these worries are exaggerated.

4.4.1. ***White***

In the sixth chapter of his book *The Manipulation of Choice* (2013), White offers a moral claim opposite to mine – that nudging by governments is by far more objectionable than the behavioral interferences of marketers. Here I will address White's objections, which I believe can be topically divided into two categories, those relating to: 1.) *the attitudes of nudgers*, and 2.) *the expectations of nudgees*.

It should be noted that White's reasons for favoring market influences are an extension of his prior argumentation against government influences. At the center of his objections is the so-called *knowledge problem*, commonly raised in nudge debates to signal the inability of government to track the settled preferences of citizens, and, thus, to nudge them towards options they would choose by their own lights. Since government lacks knowledge about what the goals and values of citizens are, any government nudging that is claimed to be motivated by subjective well-being has to be presumptuous about the content of these goals and values. This is why, White believes, government nudging is underpinned by a disrespectful attitude.

White's arguments challenge libertarian paternalism, and not a more careful pro-nudge position like my own. A democratized system of government nudging which operates in conditions of watchfulness, defended in Chapter 3, should be able to defuse the force of the knowledge problem, since it creates circumstances for individuals to consent to, or reject, the government influences that guide their behavior. Such a system safeguards personal autonomy and opens up

[94] The second question is not entirely neglected, but is mostly discussed in terms of a single regulatory mechanism – government nudges that would counteract the bad influence.

nudging to a democratic forum, which signifies a far more respectful attitude than the one White ascribes to governments under LP. However, while my account of government nudging defuses the knowledge problem, it does not fully eliminate it. Governments may increase watchfulness to make sure nudges can be circumvented, but citizens will not always be successful. Nudges can be discussed at a democratic forum, but the communication of goals and values will often be incomplete and imprecise. A watered-down version of the knowledge problem will persist.

So, let us grant to White that governments will do at least *some* tenuous (and possibly inaccurate) approximations of goals and values. White seems to believe making such approximations is offensive, or at least more offensive than how businesses operate:

> "Although the government, when acting paternalistically, presumes to promote your interests, businesses have their own interests–primarily, to maximize profit. Businesses are interested in your interests only insofar as those interests lead you to buy their products, not for your good. Whatever you think of the profit motive, the advantage of the single-minded purpose behind much business behavior is that they're not presuming to make decisions for their customers in their own interests." (ibid., 108-109).

Assume that White's claim is always true – governments always nudge in line with (an approximation of) the nudged person's background interests, while marketers nudge without regard for background interests and for their own benefit. What would explain the disrespect and objectionability of the first attitude, compared to the second? White is not explicit about this. His claim is that the first attitude is disrespectful because it is presumptuous about the nudgee's values. But would it follow that the second attitude is respectful, because it is *not* presumptuous? That is surely insufficient, for attitudes could have other properties that could signal a lack of respect.

The marketer's attitude is characterized by an assumed disregard for his target's autonomous goals. It is plausible to suggest, I contend, that his attitude is exploitative, since it shows an objectionable indifference to the value of other citizens' autonomous pursuits compared

to his own, and the readiness to use non-reflective influences to gain at their expense. More clearly than with other kinds of nudging, the attitude of marketers captures the Kantian notion of using others as means, rather than as ends in themselves. More particularly, it exemplifies the version of the means principle which condemns agents who intentionally manipulate others for their own purposes (Tadros 2011, 140). This is because the marketer undertakes the exploitation of heuristics for the fulfilment of his own ends, and because he is not concerned about whether the targets are given a chance to opt out. That seems to be a deeply disrespectful attitude. So, assuming that both attitudes are disrespectful, how do they fare comparatively? This is difficult to say, possibly because the attitudes showcase different kinds of disrespect that lack a yardstick against which they could be measured. Hence, the kinds of disrespect here could be incommensurable. Still, do the attitudes have redeeming qualities that would render them permissible? Intuitively at least, if the government is making genuine attempts at gauging its citizens' goals to help their being pursued, and granting citizens opportunities to reject the help, this attitude seems by far more respectful than the employment of tricks with a complete indifference to matters other than personal gain.

Another suggestion might be that disrespect rests with the role governments, compared to marketers, play in the social sphere. With regard to nudging, White says "we expect businesses to do it, *but we expect more from our government*", and that "we don't like to think ourselves as being at odds with our government" (White 2013, 110-111; emphasis in original). In Chapter 3, I presented my case for why a nudging government need not be at odds with its citizens, and I laid out the normative safeguards that ought to be implemented for nudging to be a collaborative venture between government and citizens. Here, I want to focus on what we can expect from marketers. On this matter, White states:

> "We know businesses do whatever they can–hopefully within the law and other ethical norms of their industry–to make money, and we fully *expect* them to do whatever they can to make money, even if we're not aware of what exactly they do. It doesn't take an extreme belief in caveat emptor ("let the buyer beware") to realize that consumers should expect businesses to manipulate their behavior to some degree and to guard themselves against it to whatever extent they can." (ibid., 109-110; emphasis in original)

Expectations of consumers seem to be relevant for, on one hand, the preservation of autonomy in the face of interferences that could divert them from pursuing their goals, and on the other, a general normative explication about the relationship between sellers and consumers. I will take each in turn. On the first count, an expectation that we will be nudged in the market could help us fend off some autonomy-undermining influences. But this general expectation is not sufficient for protecting autonomy, for it only amounts to what Bovens has called a *type interference transparency*. Bovens has stated, and I have supported his argument in Chapter 3, that type transparency alone is normatively insufficient. The mere knowledge that marketers will nudge us tells us very little about what kinds of techniques they will use (as White mentions himself), in what way they are deceptive, how numerous the influences will be, whether they are resistible, etc.

On the second count, White seems to be assuming that sellers and consumers are "at odds", by virtue of the practices used. This much is true in our non-ideal circumstances. If marketers miss out on the opportunity to nudge, they might arrive at a disadvantage compared to their competitors. In this day and age, it would seem far too demanding to pose such moral requirements to marketers. I argue for this more extensively in 4.4.2.1. We might ask, however, whether White believes sellers and consumers are at odds as a matter of necessity. If he does not, then non-ideal theorizing should give some guidance in drawing us closer to ideal circumstances. How exactly the seller-consumer relationship should take shape with regard to behavioral influences in ideal theory is not clear. But to fulfill this task, it hardly seems sufficient to look only at our present expectations about the behavior of marketers and the techniques they use in the current market setting. It is, of course,

entirely possible that White only has non-ideal circumstances in mind. However, even in these circumstances, we could conceive of factors that change the character of the relationship, such as a market culture with a high level of trust between sellers and buyers in small communities.

Lastly, a greater worry about citizens being at odds with government than with businesses might stem from slightly ideologized notions of government and business. On these notions, governments are far more powerful than businesses, and, thus, their influences pose far greater threats to autonomous citizens. Governments are often viewed as huge monolithic and monopolistic structures of power that go unchecked when they non-reflectively influence, whereas market agents are viewed as much smaller, and their influences scattered and contested by the competition. This perspective is outdated. Particularly with the rise of new informational technologies, massive companies like Google and Facebook (as well as their competitors) tower over many governments of the world in terms of power and resources, not to mention capacities to steer decision-making. And even if Internet-based companies do not have a tremendous impact on your life, it is still far more likely that, on a daily basis, you make far more choices that are steered by extremely powerful market agents, like a bank, an oil supplier, or a supermarket chain, than by governmental agencies. Thus, if the fear about being at odds with government as opposed to businesses only concerns power relations, it is certainly unfounded.[95]

4.4.2. ***Further objections***

4.4.2.1. *Demandingness*

It might be thought that to respond in the affirmative to either the ethical or the political question would entail hefty requirements for marketers (regarding the ethical question), or the

[95] I thank Andrés Moles for raising this point.

institutions that would enforce the regulation (regarding the political question). I will take each worry in turn. Is it overly demanding to require individuals to refrain from behaviorally influencing prospective consumers for profit-maximizing purposes? Thaler and Sunstein do implore the private and public sectors to refrain from these kinds of activities: "[…] we argue for self-conscious efforts, by institutions in the private sector and also by government, to steer people's choices in directions that will improve their lives" (2008, 5). As reported by Della Bradshaw, this is in line with how Thaler signs *Nudge* ('Nudge for good'), but considers it to be "a plea, not an expectation" (2015). Thaler and Sunstein seem to consider that 'nudging for good' should be a moral maxim, but not a particularly strong one. In fact, Sunstein himself points to common overriding circumstances:

> "In identifiable cases, those who do *not* exploit human errors will be seriously punished by market forces, simply because their competitors are profiting from doing so. Credit markets provide many sad examples. Consider cell phone plans, credit card plans, checking accounts, and mortgages, which usually have many good features but which are often unfathomably complex, and which can hide potentially damaging terms (such as high fees for "overdraft protection"). In all of these areas, it is possible that companies that provide clear, simple products would do poorly in the marketplace, because they are not taking advantage of people's propensity to blunder" (2015b, 10-11; emphasis in original).

Thus, in certain areas of the market at least, requiring marketers to refrain from for-profit influencing would be tantamount to asking them to allow their businesses to go under. While there might be areas of the market in which there are incentives for such restraint (for instance, preserving the company's reputation), Akerlof and Shiller state that:

> "unregulated free markets rarely reward […] heroism, of those who restrain themselves from taking advantage of customers' psychological or informational weaknesses. Because of competitive pressures, managers who restrain themselves in this way tend to be replaced by others with fewer moral qualms" (2015, xii).

Hence, it does seem to be overly demanding to pose the ethical requirement (i.e. answer the ethical question in the affirmative) in behaviorally unregulated, or scarcely regulated (i.e. non-

ideal) markets like our own. Even if we insisted that marketers act wrongly, we would likely excuse it. As Akerlof and Shiller seem to suggest, markets are an unforgiving arena of differential advantage, where staying in the race often demands employing the entire toolset of legally available means. The question about demandingness might resurface in a more developed culture of restraint. If, in a near future, a political community becomes acutely sensitive to the exploitation of bias for differential gain in markets, say, because it brings about costly negative externalities, then we might hope that marketers will be more likely to respect pleas of restraint.

The other, *political* worry about demandingness, concerns how market-restricting regulation is expected to cover the full range of behavioral market activities, and how large an institutional scheme is supposed to be to curtail it. This worry is not nearly as dire. Advocating institutional schemes for reducing the exploitation of behavioral blunders would not require the market to be regulated *across the board*. For example, counteracting the effects of a market nudge with a government nudge could be one instance of market-restricting regulation. It would not count against it that not all market nudges are counteracted. In fact, a limit to government interference could be desirable. First, governments would be stifling marketers if they were interfering into every segment of their business operations. Second, such a strategy would undermine effective watchfulness given the stacks of nudges that would thus have to be introduced (see Chapter 3). Hence, a regulation strategy to counter market nudging and protect autonomy could be much less encompassing. Its operations might be reduced to bracketing the most autonomy-threatening aspects of market interactions in areas of decision-making that are most commonly regarded as non-trivial, and counteracting them with transparent government nudges (or some other, perhaps more viable strategy described in 4.5.).

4.4.2.2. *Conceivability*

I mentioned in 4.2. that it is hard to conceive a market designed to effectively undercut non-reflective influences. What kind of form would regulation have to take to seriously curb the host of market influences? And what would the market even look like once it has been properly "purified"? Practices of advertising and other market interactions have become so hardwired into our practical conception of the market that it is difficult to envisage the end result. In short, we are not sure *how* to purge the market from bad influences, and even less whether we would find the outcome desirable.[96]

The conceivability problem mostly raises issues which we could leave to policy makers, but some normative concerns pop up as well. For instance, some may view certain aspects of barter and trade in a romantic light, and would object if the future market did not have these practices preserved. They might try to draw on arguments like that of G.A. Cohen, who claims that we have good normative reasons to conserve things that are valuable, even in the face of things that have greater value (2013, 149). In this debate specifically, if it could be shown that certain market practices involving non-reflective influences have value in terms of culture or tradition, then we might consider placing some constraints on autonomy. The burden of proof for the conservative argument here is even lightened given my non-perfectionist commitment only to *enabling* autonomy in the midst of non-reflective influences, and not to its maximization. But if the conservative argument is truly inspired by an account like Cohen's, then it would have to prove that preserving market practices playing on non-reflective influences does not also preserve injustice, which Cohen's conception of conservatism emphatically rejects (ibid., 144). I mentioned

[96] What might help conceivability is the fact that tobacco advertising has long been stifled. I thank Andrés Moles for pointing this out to me. However, we might still have trouble imagining advertisement bans across the board.

earlier that given our expectation about salespersons being far more versed in the workings of behavioral techniques than consumers, allowing the unfettered utilization of market nudges often gives salespersons a decisive bargaining power over consumers. The preservation of such distinct bargaining positions, I judged, should hardly be considered just.

The other normative issue is that if the market is hardly conceivable without market nudges, then market nudges might be a *constitutive* part of market activities. In effect, any argument against market nudges would then be an argument against markets themselves. And this would, in turn, allow opponents to dismiss my arguments because, although effects of non-reflective market influences might be deeply regrettable, they are part of an efficient system of economic coordination for which, some authors claim, we are still lacking viable alternatives.[97] This objection might very well be warranted if the reformed market would lack the drivers of behavior that we associate with economic productivity. My wish here, however, is not to mount an all-out attack on markets on the shoulders of a much more constrained criticism of market nudges. Imagine that some market nudges were shown to be inevitable, or that they were constitutive parts of practices that contribute to mutual advantage to an extent much greater than any other system of economic coordination could secure. This would, for the time being, be sufficient to defend the legal presence of these market nudges. But we should think that some forms of advertising or bartering practices that are used to swindle buyers are not constitutive, and that markets would be conceivable without them. As long as we are able to identify such practices and devise regulatory counters for them that do not undermine the market as a whole, we should be able to mount criticisms of market nudges that are independent of criticisms of markets.

[97] See, for example, Friedman (1982). For the possibilities and obstacles of making intersystemic efficiency comparisons, see Buchanan (1985, 36-46).

In fact, if market nudges could be successfully disintegrated from the market's core, then any independent case in favor of the market should find markets with fewer market nudges more desirable and more easily justifiable. For those who favor markets on consequentalist grounds, markets are justified because they facilitate mutually advantageous exchanges that culminate in Pareto optimality, an equilibrium state in which it is impossible to redistribute resources without making at least one participant worse-off. Allen Buchanan stresses, however, that attaining Pareto optimality depends upon several conditions (1985, 14-15). For the fulfillment of one particular condition – the rationality of individuals (in that their preferences are in transitive ordering) – market nudges represent an undermining force. They will often be used to distract individuals from the commitments that they want to include in their preference ordering. Buchanan points out that these are particularly idealized conditions, and that the justification of non-ideal markets depends upon how well they approximate the ideal (ibid., 15). But if market nudges detract from the correct approximation of the rationality condition, then market consequentalists would find it easier to justify settings with fewer market nudges or with regulation containing them.[98]

Non-consequentalist accounts in favor of markets argue instead that entering into market exchanges is not justified by the positive results of the transactions, but by those transactions representing an exercise of fundamental rights. The most famous right-based defense of the market, that of Robert Nozick, maintains that individual rights are respected, and that justice is

[98] One objection would suggest that Pareto optimality could still be attained in markets with rampant nudging. Strictly, a conceivable Pareto optimality is compatible with a systematic advantage of nudgers over nudgees, since benefitting the nudger will often entail worsening the nudgee. But as Sen notes, "[i]f the utility of the deprived cannot be raised without cutting into the utility of the rich, the situation can be Pareto optimal but truly awful" (1985, 10). Consequentalism is then highly objectionable if it does not discriminate between the desirable and the 'truly awful' Pareto optimalities, yet it is likely that consequentalists would look for ways to mark the latter outcomes as not properly mutually advantageous.
Another objection follows Buchanan, who claims that "competition in nonideal markets generates incentives for behavior that tends toward the more perfect satisfaction of the condition of the ideal market" (1985, 15-16). However, this claim is questionable if agents are required to utilize market nudges in order to stay competitive, as Sunstein argues (see 4.4.2.1.).

preserved, when the distribution of holdings arises from just initial acquisitions and voluntary transfers. Central to my purposes here, standard cases of legitimate transfers in holdings for Nozick are voluntary exchanges and gift-giving, while he takes fraud to be on the opposite side of the spectrum (1974, 150). But for Nozick, who deems the matter of voluntary transfer a 'complicated truth' (ibid.), it is not clear where transactions that are influenced by market nudges lie between 'voluntary exchanges' and 'fraud'. Non-consequentalists would then have a reason to oppose market nudges. The voluntariness of transfers seems to be more easily preserved when market nudges are at least minimized.

4.4.2.3. *The appropriate sites of moral assessment and institutional regulation*

Authors from the nudge debate might be reluctant to deal with market nudges because we are not settled on what the proper *site* for morally assessing behavioral influence is, and whether, in different sites, such assessment bears implications for institutional regulation. Let me explain. So far in the literature, the site for the moral assessment of behavioral influences has been the relationship between government and citizens. In this chapter, I have argued that we should extend our assessment to the site where marketers are the nudgers, and citizens the nudgees. Do these separate moral assessments influence each other? For instance, if market nudges are permissible, does this bear relevance for the permissibility of government nudges? Or, if behavioral influences are permissible in personal relationships (among friends, families, and spouses), does this bear relevance for the permissibility of market nudges?[99] Resolving these matters will hint at answers to both the ethical and political questions from the beginning of the section. Consider the following quote by Sunstein:

[99] As we will see in 4.5., the connection between the permissibility of government nudges and the permissibility of market nudges has already been established, given that government nudges are often proposed as a counteracting strategy against market nudges.

> "Much of modern advertising is directed at System 1, with attractive people, bold colors, and distinctive aesthetics. (Consider advertisements for Viagra.) Cell phone companies, restaurants, and clothing stores use music and colors in a way that is designed to »frame« products in a distinctive manner. Doctors, friends, and family members (including spouses) often do something quite similar. Is romance an exercise in manipulation? Maybe so. Is medical care? Is the use of social media? A great deal of conduct, however familiar, can be counted as manipulative in some relevant sense; but it would be extreme to condemn it for that reason." (2016a, 60)

Sunstein's point seems to be that if we were to highlight manipulative tendencies across the vast areas of human interaction, the wrongdoing of many single instances of influencing would be downplayed and would not warrant expressions of condemnation. In short, Sunstein suggests that market influences are not so worrying given the pervasiveness of influences in other areas. This is particularly so, Sunstein seems to suggest, by virtue of manipulative influences being entrenched in personal relationships:

> "In ordinary life, we would not be likely to accuse our friends or loved ones of manipulation if they characterized one approach as favored by most members of our peer group, or if they emphasized the losses that might accompany an alternative that they abhor, or if they accompanied a description of one option with a frown and another with a smile." (ibid., 63)

I disagree with Sunstein here. If we are manipulated in the private domain, in ways in which our manipulators deliberately aim to non-transparently trigger our heuristics, and especially if these influences push against our deep commitments, we are justified in raising complaints. Our influencers often know full well when we would rather engage in matters reflectively or that they are triggering us to act in ways that are incoherent with our settled preferences. Influencing between close friends, family, and partners might sometimes be tolerated and even justified, but this is mostly because the nature of these relationships is such that these people are familiar with our goals and can be allowed to act on our behalf. Certain practices of influencing are also permissible given the specific character of social interaction. Consider the example by McCrudden and King:

> "Suppose that an attractive man flirts with a friend in order to get her to join him for dinner. He is aware that in using charm, she is more likely to join him. In our view, there is nothing manipulative in this exchange provided that the woman is aware of the ploy – she sees it and (perhaps) delights in acquiescence […] Suppose by contrast, the man flirts with the woman for an end that he conceals from her. In such a case, this *would* be manipulation because he uses her to obtain an end in a manner that subverts and is non-cognizable by her own rational thought processes." (2016, 116; emphasis in original)

I agree with McCrudden and King's case here. Thus, I find it plausible that the moral assessment of behavioral influences is suitable across all three mentioned sites – "personal nudging", market nudging, and government nudging. But let's turn back to the political question now. Which of these sites are suitable for institutional or other social regulation? I respond here by drawing inspiration from the seasoned debate on the *site of distributive justice*. This debate starts from Rawls's suggestion that principles of justice primarily apply to what he terms 'the basic structure'. According to Rawls, the basic structure refers to major social institutions, which include "the political constitution and the principal economic and social arrangements" (1999a, 6), and should not be "confused with the principles which apply to individuals and their actions in particular circumstances" (ibid., 47). The purpose of the basic structure, according to Rawls, is "to secure just background conditions against which the actions of individuals and associations take place" (1996, 266). Rawls's institutions, the realizations of systems of rules, specifically include markets (1999a, 48). The inclusion of markets into the basic structure is particularly useful here – it shows why market and government nudging is an appropriate site for regulation, whereas "personal nudging" is not.

Can the prevention of market nudges and their effects be derived from the basic structure? On a standard understanding, the basic structure refers not to the immediate laws and regulations, but to a more general level of institutional consolidation. Still, Rawls seems to think that the basic structure "enforces through the legal system another set of rules that govern transactions and

agreements between individuals and associations," which include transgressions such as "fraud and duress" (1996, 268). It could be plausibly stated, I believe, that the exploitation of cognitive heuristics in market interactions erodes the background conditions of justice. For instance, the systematic exploitation would on average favor the sellers over the buyers, undermine autonomy for the purposes of differential advantage, and violate the difference principle. If this would indeed be the case, we would have at least *pro tanto* reasons to derive legal rules from the basic structure that would remedy the resulting deviations from justice.

In advocating that regulations be derived from the basic structure, Rawls is also acutely aware of the other two problems that I have discussed – demandingness and conceivability. The rules that would guide individual behavior in market transactions must be "feasible and practicable," and "cannot be too complex, or require too much information to be correctly applied" (ibid., 267), lest they become "an excessive if not an impossible burden" (ibid., 266). So, in order for rules of market interaction to be successfully derived from the basic structure, they will need to pass a *practicability test*. I will not be able to lay out the exact blueprint for practicable regulation here, but I will offer some useful examples in 4.5. that I believe could beat the test.[100]

[100] That Rawls would agree with my assessment that many market practices should be regulated is confirmed by his take on the social status of the liberty to advertise:

> "advertising tries to influence consumers' preferences by presenting the firm as trustworthy through the use of slogans, eye-catching photographs, and so on, all designed to form or to strengthen the habit of buying the firm's products. Much of this kind of advertising is socially wasteful, and a well-ordered society that tries to preserve competition and to remove market imperfections would seek reasonable ways to limit it. The funds now devoted to advertising can be released for investment or for other useful social ends. Thus, the legislature might, for example, encourage agreements among firms to limit expenditures on this kind of advertising through taxes and by enforcing such contracts as legally valid." (Rawls 1996, 365)

4.5. Policy applications

Now that I have established that the market setting is an appropriate site for regulatory intervention, I want to conclude the chapter by discussing several policy applications that could alleviate the threat that market nudges pose to personal autonomy (but, possibly, to other values as well). I should reiterate that the consideration of policy applications is only aimed at reducing market nudging, not rejecting the market or revolutionizing it to the point of seriously undermining the incentives that drive its efficiency. We should take care not to throw the baby out with the bathwater. Many instances of intended non-reflective influence are deeply integrated into market interactions (such as product placement) and it is difficult to conceive the market setting wholly without them. The proposed policies are only meant to help individuals more effectively pursue their conceptions of the good in the face of detrimental influences.

Some policy results would likely be accomplished with the institutional incorporation of the principle of watchfulness, which I have required in Chapter 3 as the condition to make nudges appropriately transparent. Watchfulness would give citizens a basic sense of how non-reflective influences work, how significant they are in steering behavior, and how they could undermine or aid their autonomous pursuits. To citizens who want to be on the lookout for any and all market nudges, watchfulness provides a useful tutorial for developing their skills of noticing influences. But unlike the program of government nudging, in which watchful citizens can track and micro-manage nudges with the help of government resources, and which is tailored as a cooperative venture between government, experts, and citizens, the area of market nudging is not conceived as user-friendly. To successfully navigate the landscape of market nudges would require much greater efforts from citizens (to build up active awareness of market nudges that surround them) and from governments (to provide citizens with as much information to do so). But this requirement suffers

from all the objections posed to the maximalist conception of watchfulness presented in 3.5.2. I thus turn to other complementary policy solutions.

The first possible strategy is to use government nudges to counteract market nudges. Instead of only utilizing nudges to steer the uncontroversial goals of the majority, governments should apply them in cases where marketers influence choices with big stakes, pertaining to health, finance, big purchases (accommodation or car), etc. According to Blumenthal-Barby:

> "Much of our choice environment is structured by people who have no concern for our well-being, e.g., advertisers who subtly, creatively, and pervasively prompt us to buy their products, consume their goods, and adopt their way of life. In light of this, it would be preferable to have choice architects concerned with structuring choices reflectively and responsibly in a way that makes people better off even if they are occasionally subject to biases and errors in judgment and decision-making." (2013, 185)

Despite mostly focusing on consumer tendencies rather than on the marketer's tricks, Thaler and Sunstein also seem to be sympathetic to such a conception when they discuss credit brokers and mention other walks of life in which nudging mainly benefits the nudger (2008, 135, 239). As Sunstein notes:

> "If government is targeting System 1 […] it may be responding to the fact that System 1 has already been targeted, and to people's detriment. In the context of cigarettes, for example, it is plausible to say that a range of manipulations – including advertising and social norms – have influenced people to become smokers. If this is so, perhaps we can say that public officials are permitted to meet fire with fire." (Sunstein 2016a, 63)

A program of watchful nudging, as I already argued, should be driven by considerations of where and when market nudges can undermine autonomous choice the most; transparent nudging could be incorporated to counteract these effects. However, this means has limited potential. It can only counteract some market nudges, lest we risk undermining watchfulness due to government nudge stacking. If counteracting was the only policy means to counter market nudges, but with watchfulness in place, then it would have little effect against market nudge stacking. We would

end up having to choose between government nudge stacking, on one side, and market nudge stacking, on the other. Another difficulty is that marketers are better placed to quickly and effectively employ new market nudges, and thus counteract the counteracting of government. The counteracting strategy is, thus, permissible and desirable, but insufficient.

A further method to consider incorporating into the policy package is Adam Oliver's 'budge' strategy (2013), which focuses on providing public officials with "an education in behavioural economic concepts" and "offering [them] potential insights into where and how their citizens' cognitive limitations are being exploited excessively" (ibid., 698). The ultimate goal of the budge strategy is to hamper the "profit-oriented industry" by informing regulators' decisions "on where and how to regulate (for instance, traffic light food labelling), and [to] ensure that public officials gain a better understanding of their own decision-making limitations" (ibid., 698-699). I agree with Oliver. A further requirement from the informed public officials, however, would be to come up with concrete policy that would at least partly mitigate market nudge stacking. I speculate on two kinds of strategies here.

The first line of strategies relates to the control of *content*. First, some advertisements nowadays hold no actual content about the product or service that they are selling. Old advertisements were conceived as providing information about products, while nowadays they often fail to do so entirely. It would not be too controversial to argue that such ads should be disallowed, or that some regulation on minimal content should be required, in order to reduce distractions from relevant information. Second, I mentioned Wendel's point in 4.2. that marketing companies are often under nondisclosure agreements about their work. For similar reasons as with

the first regulation, a different regulation might require such firms to reveal their methods.[101] Third, it is not perfectly clear nowadays whether pieces of content in public spaces are advertisements or not. One common occurrence in recent years is commercials disguised as news. Since consumers often fail to spot the difference between news and commercials, and thus fail to be on the lookout for market nudges, regulation could require that news pieces and commercials be clearly marked and distinguished from one another.

The second line of strategies relates to the control of *public spaces*. Consider the 'Clean City' law, which was enforced from 2007 in the city of São Paulo to remove all advertisements from public spaces, including buildings, posters, billboards, buses, and taxis; even handing out pamphlets. Such a policy solution seems to directly relate to market nudge stacking, which is exacerbated with the extension of public communication channels. It significantly alleviates the extent to which consumers are exposed to market techniques. Of course, a policy solution need not go as far as the Clean City law. We may sometimes be satisfied with organizing advertisement-free city zones. Our efforts of minimizing the detriments of market nudging should be counterbalanced with possible bad side-effects to markets that policies might cause. But we need not think of the proposal as too radical or ruinous to market efficiency. Consider that our virtual spaces can already be purged of advertisements with the help of software – so-called *adblocks* – and that web marketers can ask users to opt out.

Some regulations will relate to both content and space. For instance, a famous field experiment by Robert Zajonc and D.W. Rajecki shows that the positive attitudes towards stimuli can be enhanced merely by repeated exposure (1969). This finding easily explains why so many

[101] I owe both policy ideas to Andrés Moles.

marketers put so much importance on public presence. In response, public officials could introduce regulations that limit the extent to which a particular company can use public spaces for advertising. Sometimes, we might even believe that outright bans are justified. For instance, the European Commission introduced a ban on all tobacco advertising in print, radio, online, or in sponsor events (2003).

All of these regulations concerned advertising. Further policy ideas could relate to other aspects of market interactions, such as the regulation of market spaces or pricing. I leave these to policy makers, and return to the theoretical foundations of behavioral enhancement that concern other-regarding consideration, which I tackle in the next chapter.

Chapter 5: Moral Nudges

In the previous chapters, I have made efforts to broaden an overly narrow moral debate on behavioral influences. Yet, in one significant respect, my ethical treatment has mostly remained loyal to the source, insofar as the permissibility of nudging has been assessed in terms of *self-regarding* considerations. Specifically, I was testing whether nudges and other behavioral influences are permissible in liberal democracies almost exclusively in terms of *individual* concerns. As noted in the Introduction, this approach has become the norm among moral philosophers interested in nudging, due to the nudge project's scientific background in behavioral economics, where welfarism reigns supreme. In a similar vein, philosophers have predominantly observed nudging through the lens of paternalism, mostly because of Thaler and Sunstein's commitment to 'libertarian paternalism'. However, as I hinted in the Introduction, the umbilical cord between welfarism and LP on one side, and nudging on the other, is only implicated and rarely argued for.

The permissibility of behavioral influences can, instead, be assessed in terms of how the influenced behavior of individuals affects others. The upshot of quite a number of nudges is expressed in this way. Take the famous urinal fly: Karen Yeung argues that the benefits from men's improved aim accrue not to themselves, but to those who will later use the urinal, or those who are tasked with cleaning it (2012, 124). Similarly, in an elicited example of nudging – shifting the default in organ donation schemes – the advantages for the donors are clearly secondary to the benefits for those receiving the organs. Additionally, some nudges standardly discussed in self-regarding terms can be given plausible other-regarding justifications. Cafeteria arrangements, for instance, could help individuals in avoiding fatty foods, but another attempt to justify this nudge could be to stress its effect on the reduction of health care costs.

This realization faces us with an important moral question – is it permissible to influence people's choices for the sake of others, and specifically, to facilitate the discharging of their moral duties? Empirical research about the potential of nudges on driving, tax abidance, organ and blood donation, charity giving, and environmental protection has produced encouraging results, but moral philosophers have so far done little to highlight the important moral conundrums. This concerns not only the ethicists interested in nudging, but also those working on *moral enhancement*, which particularly evaluates biochemical methods of moral improvement. These authors have seldom made note of behavioral influences, let alone compared the moral properties of the two types of influences.[102] This is odd, given that 1.) there are important similarities in the conceptual frameworks of the two projects; 2.) they build on a similar psychological literature that points to causes of moral misdemeanor; and 3.) the most common objections against bioenhancements can also be raised against morally-minded nudges. Throughout, and especially in 5.3., I show that many of the ethical worries in our analysis of *moral nudges* can be adopted from the moral bioenhancement debate.

In 3.4., I suggested that making nudges transparent, and thus giving nudgees a chance to autonomously navigate the landscape of government influences, can take a back seat when stakes are high and other values take precedence. The enabling of autonomy can give way to other considerations when nudges can facilitate the discharging of important enforceable and non-enforceable duties. The moral nudges that I will end up defending in this chapter are an extension to my commitments to non-perfectionist political liberalism, and will strictly be used for non-

[102] Notable exceptions are Pugh (2017, 84), who states that any objection to moral bioenhancements must show how they threaten freedom as opposed to externally imposed environmental influences (like nudges), Douglas (2018), who claims that it is hard to find any property that would explain why moral intuitions decisively favor nudges, but disfavor bioenhancements, and Bublitz and Merkel (2014, 69), who believe persons have more control over nudges insofar as they are perceptually processed.

perfectionist political considerations, such as the prevention of some grave harm. This chapter spells out an account of the permissible use of such moral nudges. Given that personal autonomy gives way, I claim that the constraints on moral nudges are much weaker than those on self-regarding nudges by governments. When autonomy is decisively outweighed by conflicting values, moral nudges do not require transparency of the democratic pedigree that I earlier defended for self-regarding nudges.

On the flip side, however, non-transparent moral nudges will have the capacity to produce some deeply regrettable side-effects. Firstly, because such nudges result from bypassing reasoning and jumpstarting motivation, they will often undermine responsiveness to moral reasoning that should drive the discharging of moral duties. As a result of moral nudges, nudgees will often merely *conform* to moral duties, rather than *comply* with them.[103] Secondly, because moral nudges prompt individuals without offering reasons, they could induce uniformity about values, thus stifling democratic deliberation about important collective duties. Both side-effects could deeply reflect on our moral power of understanding, applying, and acting from a public conception of justice, or the way we may hope to realize it. I will show, however, that while these side-effects are conceivable, and could be regrettable, they emanate only *contingently* from the use of moral nudges. In many cases, moral nudges will make compliance and deliberation no worse than in the choice contexts that are altered by them.

The chapter proceeds as follows. In the next section, I specify the influence pervasiveness that contributes to moral agents not living up to their moral duties, and I outline the behavioral influences that are available for countering their effects. Second, I argue for the permissibility of

[103] See Raz (1999, 178-182).

moral nudges in cases of guiding people towards discharging their enforceable and at least some non-enforceable duties, but I also describe types of moral uncertainty, some of which will allow nudging, but with certain constraints. Third, I compare the conceptual frameworks of the debates on moral bioenhancement and nudging. Here I explore which moral worries and objections from the moral bioenhancement debate are applicable for assessing moral nudges. I explain why moral nudges could threaten moral compliance and democratic deliberation, but show that this would occur only contingently. I further elaborate in the fourth section how previous arguments in the dissertation strengthen the case for moral nudges. I discuss moral nudges in cases where influences are inevitable and their effects predictable, where they economize cognitive resources, and where they are reasonably expected to enable personal autonomy on a resource-based account that I have defended.

5.1. Behavioral influences on moral abidance

5.1.1. ***Detrimental influences***

In their book, *Unfit for the Future* (2012), Ingmar Persson and Julian Savulescu argue that the evolution of our moral psychology has been lagging significantly behind our great strides in technological development. Technology has given humans numerous tools to cause great harm. Almost any person in industrialized societies can now wreak havoc by doing something as simple as driving a car through a crowd. Simultaneously, there are fewer occasions for people to benefit others to such a great extent (ibid., 12-13). Furthermore, most contemporary problems, such as climate change, global poverty, or various kinds of group-based discrimination, are of a global ilk. Persson and Savulescu suggest that our inability to deal with these problems is owed greatly to our

psychological inadequacies, blocking us from behaving altruistically and acting from a sense of justice.

Recall from the Introduction that my account of a behaviorally-enhanced liberal democracy would be driven by the enabling of two Rawlsian moral powers, one of which is the capacity to understand, apply, and act from the public conception of justice. Burdens of moral psychology, if they can be lifted, would require that we form our institutions to overcome these disruptive effects. If we, however, cannot fully relieve ourselves of these burdens, but could utilize the deficiencies of moral psychology to produce just outcomes, then perhaps this is what justice would require, rather than leaving weighty duties to be ignored. But before we take to justifying moral nudges, let me first present what these cognitive shortcomings consist in.

Persson and Savulescu argue that we are bad at dealing with global problems because we possess a "localized" morality. Having evolved in close-knit communities and having to face imminent dangers in hopes of survival caused us to be biased towards the here-and-now (ibid., 27). Our considerations of care usually extend to those near and dear to us, and not too far past the immediate future. Thus, our capacity for care is expended on those who we perceive to be members of the in-group, while we remain distrustful of strangers. Exposure can boost in us a "strong sympathy or compassion for beings who suffer before our eyes", yet we are rarely moved by the plights of "strangers, distant in space or time" (2019, 8).[104]

[104] To be sure, individuals are sometimes justified not to take up collective duties knowing that psychological shortcomings will hinder them and their peers from doing their part. For some duties at least, we are not obliged to do our part if we have good reasons to believe others will fail to do theirs. In this chapter, I mainly discuss cases where some collective duties are widely acknowledged, or where individuals partaking in the duties express a sincere intention to do their part, but fail to do so owing to psychological shortcomings and ill-suited choice environments. My case in favor of moral nudging aims to secure the conditions in which the fulfilment of collective duties becomes possible. I thank Andrés Moles for pointing this out to me. See also the following chapter (6.3.), where I argue that a public commitment to some other-regarding nudge project can facilitate conditional cooperation among citizens.

Furthermore, the compassion that we are able to muster is not responsive to the rising numbers of people suffering, making us unable to adequately respond when such numbers multiply (2012, 30, 62-63).[105] The distrustful environment of the global community is in fact among the primary factors for why global problems bloom. Individuals are hindered by a group bias to take on a global problem with distant strangers, which leads to collective action deadlocks in which problems are amplified (ibid., 69-70). In such cases, individuals feel little responsibility for the problems caused. This is because, on one hand, the harm is a cumulative effect of the inaction of many, next to which the inaction of the single individual is negligible (ibid., 63). On the other hand, individuals predominantly see their global duties as those of prevention. This is unfortunate, given that our moral psychology resonates much more strongly with the wrongness of causing harm (like drowning someone) than with the wrongness of omitting to prevent harm (allowing someone to drown) (ibid., 22).[106]

Let's focus on a particular global problem – risks of environmental catastrophe. A number of further cognitive failures contribute to the inability to meet this problem head on. For instance, Kim Kamin and Jeffrey Rachlinski show that people will be far more likely to state that precautions were in order after some catastrophe, like flooding, has occurred, than before it has occurred (1995). This *hindsight bias*, they further prove, is not dispelled even if it is specifically pointed out to people and they are instructed to avoid its effect.[107] Furthermore, Daniel Kahneman and Jack Knetsch argue that if people do invest in a good cause, they do so only to reach their personal

[105] See Desvouges et al. (1993), who tested how high a tax increase people would be willing to accept for saving birds from oil spillages. The estimated group of saved birds varied between 2,000, 20,000, and 200,000, but there were hardly any differences in the amounts suggested by the respondents.
[106] This aspect of our moral psychology could be said to ground intuitions to morally differentiate acts and omissions, which I have argued against in some cases of nudging in 2.4.2.
[107] This last point will prove relevant later when we discuss the possibility that, in some cases, individuals will not be able to overcome non-reflective biases when reasoning about them. It should be permissible to employ nudges without the concern of enabling moral compliance in such cases.

'moral satisfaction' threshold, which is often divorced from a meaningful contribution to the cause. Individuals are also overly optimistic about resolving collective problems. They might exhibit a *planning fallacy*, showing unfounded confidence about how long it will take them to perform a particular task (Buehler et al. 1994), or they might take for granted that some unforeseen technological fix will occur in the future (Persson and Savulescu 2012, 74). Eliezer Yudkowsky points out still further biases in reasoning:

> "[People] may focus on overly specific scenarios for the future, to the exclusion of all others. They may not recall any past extinction events in memory. They may overestimate the predictability of the past and hence underestimate the surprise of the future. […] They may be contaminated by movies, where the world ends up being saved […] [T]he extremely unpleasant prospect of human extinction may spur them to seek arguments that humanity will *not* go extinct, without an equally frantic search for reasons why we *would*." (2008, 110-111; emphasis in original)

5.1.2. ***Moral belief and moral motivation***

The cognitive biases described can often prevent individuals from forming and endorsing moral beliefs that are thought to motivate them to discharge their duties. One such belief, for instance, might be that we ought to keep our carbon footprint in check by reducing our fuel consumption. But it is also possible that individuals are, more often than not, able to endorse such moral beliefs, but are hindered in generating the motivation to act on them by the aforementioned biases. This leads us to a contentious philosophical debate about the relationship between moral beliefs (believing that one ought to/has reason to φ) and moral motivation (being motivated to φ). I take a moment here to outline this discussion and show which understanding of this relationship is ruled out by a deeper appreciation of cognitive heuristics.

Internalists about moral motivation believe that motivation is internal to moral beliefs and judgments. Believing that I have a reason to φ entails, *necessarily*, according to the internalist, that I have motivation to φ. If I were to claim that I believe that I ought to φ, but showed no motivation

to φ, then the internalist would say that my expression of moral belief was insincere. Externalists about moral motivation, on the other hand, hold that moral beliefs and moral motivation come together only *contingently*. To externalists, moral motivation occurs when the reasons derived from the moral belief align with a desire to φ (Rosati 2016).

Internalists object to externalists stating that their picture of moral motivation cannot properly account for the seemingly reliable way in which moral motivations shift in line with changes in moral beliefs. It cannot be a coincidence, says the internalist, that moral beliefs steer motivation so predictably. Externalists fire back by claiming that their position is far more suited to explain some intuitive cases: 1.) individuals could judge that they have moral reasons to φ, but not generate motivation because they believe they would fail to φ; 2.) individuals could judge φ to be moral, yet still wonder why they should act morally; or 3.) individuals could be motivated by a belief in φ for a long time, but lose their motivation without changing their belief (ibid.).

Importantly for my case here, the consideration of non-reflective influences and biases should confirm that moral beliefs and moral motivation can and do come apart. It seems true for many of us that we sincerely believe we have decisive reasons to tackle problems like alleviating climate change, but find ourselves failing to do so (and then just as sincerely regretting it). The detrimental influences described in the previous subsection stand in the way of building up such motivation. In response to this objection, contemporary internalists have significantly qualified their positions. Sigrún Svavarsdottir (1999) states it is compatible with internalism that moral judgment generates some moral motivation, but that it is defeated by conflicting desires or cognitive blunders, like akrasia. But as Rosati (1999) points out, such a qualified internalism and externalism may not be overly distinct positions.

Externalism seems to gain the upper hand with the consideration of behavioral influences. Still, internalism may hold some intuitive grip over us. For instance, if A believes that he ought to φ, and has generated a motivation to φ, while B believes that he ought to φ, but hardly shows any signs of motivation to φ, then it seems at least intuitive that A's belief is more sincere than B's. However, it may still be possible to explain this from both positions. A qualified internalist would require that a φ-ist believes he ought to φ and, as a consequence, is at least slightly motivated to φ. An externalist may require that a φ-ist has a stable desire to do what he judges to be moral and that he judges φ to fit the bill, even if the desire is defeated by conflicting desires or biases. I will take it that a person's commitment to, for instance, environmentalism or poverty alleviation, is explained in one of these two terms. However, cognitive heuristics will rule out a strong version of internalism, since that position requires a much tighter connection between moral beliefs and motivations than can be found in the non-ideal world.

5.1.3. ***State-of-the-art moral nudges***

Generating motivation or overcoming behavioral hinderances is difficult for individuals not only due to a lagging moral psychology, but because of "a relevant choice architecture" (Sunstein and Reisch 2013, 402). The decision-making context might further exacerbate the difficulties of acting in line with moral reasons and duties. In some areas, relevant decision-making is already heavily influenced in directions that would hardly be favored by our intuitive moral judgments. T.J. Kasperbauer, for instance, notes that "[e]nergy consumption is massively influenced by external factors and is thus already subject to nudge-like interventions" (2017, 52). In this subsection, I list several tested behavioral techniques, by area of potential moral duty, and

conceptually configure the kinds of moral nudges that will be assessed throughout the remainder of the chapter.[108]

Let's start once more from environmental protection. Green nudges (Schubert 2016) lead the way in terms of the numbers of beneficial techniques available. Among these, green defaults seem to hold the most potential for changing outcomes. For instance, people favor by a great margin a more expensive, yet more sustainable light bulb to a less expensive and sustainable one if the household is already fitted with fixtures for the former (Dinner et al. 2011). Switching from a single-sided to a double-sided printing default at one Swedish university resulted in a 15% reduction of paper use, an effect that is not in the least rivaled by asking or educating people to contribute (Egebark and Ekström 2013).[109] Moreover, in spite of a general statement by many Germans that they would switch to green energy given the opportunity, few do so. There are cases of towns and regions where the green usage rate is over 90%, due to green defaults (Pichert and Katsikopoulos 2008).

Still, other heuristics can be triggered to produce environmentally-friendly effects. Thaler and Sunstein note that switching the frame of car fuel consumption from 'gallons-per-mile' to 'miles-per-gallon' nudges people into greener options when purchasing a car (2008, 204). Pointing out social norms to reduce energy consumption (Allcott 2011) or prompting the reuse of towels in hotels (Nolan et al. 2008) proved, once again, to be more effective methods to get people to conform than explicitly stating moral reasons for action. Such techniques, claims Schubert, are

[108] My inquiry will not question whether these methods reliably produce outcomes that the evidence seems to support. Understandably, if any of these techniques were proven ineffective or produced unanticipated side-effects that could pose significant harm, the normative implications that I later discuss would also come into question.
[109] According to the same study, the default switch is also significantly more effective than a 10% paper tax.

appealing either to the targeted individuals' desire of producing an attractive self-image, or their inclination to 'follow the herd' (2016, 7).

An oft-neglected, yet one of the most relevant areas of morally responsible action is traffic. If nudges in traffic can be proven reliable and safe, they can undoubtedly prevent countless fatalities. Possibly the most popularized traffic nudge is reducing vehicle speed at Chicago's Lake Shore Drive. In order to get drivers to slow down ahead of a series of curves, the lines on the road are far apart at first, but are then painted closer and closer together to give drivers the illusion of speeding up (Thaler and Sunstein 2008, 37-39). Taking a note from this technique, trees were planted on the side of the road approaching Norfolk City in the UK at decreasing distances, to give drivers the same impression as above. As I mentioned in 3.4., fake potholes on the road could be a particularly effective *ex post* transparent nudge, and so could fake speed bumps, or visible cameras on the side of the road (fake or otherwise).

Here I skim through some other areas in which moral nudges were successfully tested. Default schemes are particularly effective for boosting organ donation in multiple European countries. Consider Thaler and Susntein's comparison of Austria and Germany, two similar countries in sociocultural terms. Germany, in which donors have to specifically communicate their intention to donate, has a donation rate of 12%. In Austria, on the other hand, where presumed consent, i.e. an opt-out regime, is in place, 99% of citizens donate (Thaler and Sunstein 2008, 178-179). In the area of group-based discrimination, Bohnet et al. (2015) have recently proposed an 'evaluation nudge', which consists in evaluating job candidates jointly rather than separately. Their study shows that evaluators are more likely to focus on candidate performance in joint evaluations, while separate evaluations leave plenty of room for group stereotypes to kick in. With a compilation of techniques, the Behavioural Insights Team (BIT) of the UK's Cabinet Office

showed that citizens can be nudged into tax compliance, particularly by appealing to social norms and by getting them to sign their tax forms before they fill them out (2012). Finally, Michael McPherson and Matthew Smith have claimed that nudges can be utilized to promote equality, namely in political influence (2008).[110]

As could have been gathered from the listed examples and my general approach to nudging in the dissertation, I focus on heuristic-triggering nudges. I explained at length in Chapter 3 that these nudges are *ex post* transparent or fully non-transparent to an unassuming and unwatchful citizenry. But if their purpose is to facilitate the discharging of moral duties, and these duties are such that they outweigh considerations of personal autonomy, then moral nudges, much like market nudges, will deviate from the standard understanding of 'nudges' in the literature. This understanding stipulates that nudges must be mild and sensitive to agents' settled preferences. But if autonomy takes the back seat, then moral nudges are not obviously burdened by such constraints, and might take on less resistible forms to make a difference in the face of looming calamities. Even Thaler and Sunstein acknowledge that for, say, "environmental problems, gentle nudges may appear ridiculously inadequate – a bit like an effort to capture a lion with a mousetrap" (2008, 184). Hence, I will use a broader understanding of 'nudge' here to include behavioral influences that are not mild, designed to favor settled preferences, or become resistible once overt to their targets. Once again, as in previous chapters, if readers are unsatisfied with this conceptual broadening, they should think of moral nudges as 'nudges*'.

Heuristic-triggering moral nudges will have one of two effects – 1.) they will bring the nudgee's behavior into alignment with their moral beliefs (either by means of bypassing reasoning

[110] I have left out an important area – nudges for poverty alleviation. A compilation will be detailed as a special case in 6.1.

or enhancing motivation to act upon a belief), or 2.) they will steer behavior regardless of underlying beliefs. In either case, they will not be appealing to moral reasons for action. Yet, in the first case, they may have a motivational effect that will make already endorsed moral reasons more conscious, increasing moral compliance.[111] In the second case, they might be more easily resisted if agents endorse moral goals that run contrary to their effects. Moral nudging should, hence, not be conceived as purging all moral reasoning or brainwashing people into adopting moral beliefs that seem particularly alien. Granted, moral nudges could *indirectly* contribute to changes in moral beliefs, if nudgees observe, ex post, the behavior that nudges have sparked as giving them reasons to act accordingly in the future. But in such cases, moral reasoners do nothing that we would regard controversial in other cases – they pick out exemplary moral behavior that they believe is worthy of endorsement. The techniques themselves would no longer be operating 'qua nudges', but as indirectly stirring moral reasoning.[112]

A final note before pressing on to a positive case for moral nudges. Evan Selinger and Kyle Whyte have objected that nudges at best work as techno-fixes that "cannot solve complex policy problems" (2012, 26). Similarly, John et al. stated that nudges "will not be enough on their own to combat climate change" (2009, 368), and in a rare reference in the moral bioenhancement literature, Persson and Savulescu voice their skepticism about the suitability of nudges "to induce behavioural changes that should be radical and permanent" (2012, 79). My arguments in favor of moral nudges do not presuppose that they are be-all and end-all policy solutions that will single-

[111] As opposed to this, motivational effects could merely increase the conformity of nudgees and have them act on reasons that are not central to the case. Take the famous 'Don't mess with Texas' anti-littering campaign, which appealed to men's Texan pride, rather than reasons against littering (Thaler and Sunstein 2008, 60). Similarly, social norms could appeal to feelings of shame that you are not conforming to an acknowledged cause, rather than to aspects of the cause itself. More on the conformity/compliance distinction in 5.3.1.

[112] For a more direct way of using nudges to facilitate moral learning, see Engelen et al. (2018) for their case of using nudges in moral exemplar stories.

handedly solve the burning problems that give rise to moral duties. Nudges are not a silver bullet. I merely state that nudges will suffice to appreciably reduce shirking from these duties. Whether such nudges are desirable compared to the properties and effects of other policy solutions is a matter of practical consideration. Hypothetically, moral nudges could 1.) compensate for the coercive failures of institutions in non-ideal circumstances, 2.) be similarly effective, yet significantly cheaper than most alternatives, 3.) be significantly more effective than, say, raising awareness and education, or 4.) they could raise fewer ethical concerns than similarly effective options, like biochemical methods for moral enhancement.

5.2. Enforceable and non-enforceable duties

We have reasons to accept limitations on our freedom when it is beyond doubt that we owe others strong duties. It is widely acknowledged that some duties are *enforceable* – that governments can impose coercive rules so that these duties would be respected and discharged. A common example is not causing others harm, or preventing, at minimal or no cost to oneself, that others suffer harm. For liberals, harm has long been the most obvious, and for some, like John Stuart Mill, "the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will" (1859, 14).

Now imagine that government has difficulty codifying effective rules that would coerce subjects into abiding by an enforceable duty. Suppose that safety in traffic, an area with a great capacity for harm, is fraught with rules and guidelines that all participants readily acknowledge, but which fail to secure the outcome for which they are designed – that all participants conduct themselves without harming others. Suppose that this is due to a systematic attention deficit on the

part of the participants. Non-transparent and non-democratic traffic nudges, if reliable in increasing abidance, seem to be permissible methods to get drivers and pedestrians to conform. The duty to prevent harm in traffic is akin to rescuing others at very little cost to oneself, which, moral philosophers believe, standardly produces a strong moral duty. Similarly, if we could use moral nudges to get people to rescue others at little cost to themselves when such circumstances arise, then this would be permissible, given that we are getting them to perform actions already prescribed by an enforceable moral duty (Clayton and Moles 2018, 241).

I assumed in my hypothetical example that drivers readily acknowledge and endorse the rules and guidelines that are meant to ensure traffic safety. Suppose now the drivers are aware of the shortcomings of those rules, as well as the adverse effects of their attention deficits on their abidance. If so, drivers have good reason to accept covert behavioral influences that bear little or no costs to themselves, even though these influences cause them to fail to engage with rules consciously, at least while they drive. Persson and Savulescu make a similar point about accepting moral bioenhancements: "it may be quite hard for you to do [A], so hard that it is more likely than not that you will fail. If there is some means, M, that would make it easier for you to succeed in performing A, it may be that you ought morally to apply M" (2019, 7). But it is also possible in non-ideal circumstances that participants will not endorse an enforceable moral duty such as this, even if it comes at negligible cost. Still, they would have no reasonable complaint against the government if it applied covert behavioral influences in order to secure their conformity. Citizens do not have the discretion to reject their enforceable moral duties, or costless interventions that are aiming to bring them in line.

Other duties are non-enforceable. Of those, some are non-enforceable due to feasibility constraints (Moles 2015, 660), meaning that government has failed to produce a set of practicable

rules ensuring that these duties are discharged (although this may only be temporary). Let's return to environmental degradation, a global problem that tends towards producing grave, if not ultimate harm. As of yet, global and local institutions have only devised ineffective ways of slowing down the cumulative harm. Other non-enforceable duties are of a more principled kind, where interfering with the performance of one duty could be in conflict with some other duty, or would come in the way of the duty being properly discharged (ibid., 660-661). For instance, most people acknowledge that they have a duty of promise-keeping to friends. But in a great majority of cases, people owe friends not only that promises be kept, but that they generate the motivation and act on those promises without the help of others. Furthermore, it might be argued that getting people to act on their promises by means of nudges would be a serious case of government overreach into private lives.

In this chapter, I endorse the view that moral nudging is permissible for duties that are enforceable and feasibility-constrained non-enforceable. As for cases of principled non-enforceability, moral nudging will either be impermissible (as in the case above), or significantly constrained.[113] A significant point of contention will be what duties we should include on the list of enforceable and feasibility-constrained non-enforceable duties, given moral disagreement in a reasonably pluralistic liberal democracy. Arguing in favor of moral bioenhancement, David DeGrazia suggests that we should "[s]tick to improvements that represent *points of overlapping consensus among competing, reasonable moral perspectives*," including but not limited to consequentalist, deontological, and virtue-based views (2012, 364; emphasis in original).

[113] More on this last point later in this chapter and in Chapter 6.

Whatever duties are found in the overlapping set of these views should be suitable for moral bioenhancement, and if these methods are similar in relevant moral terms, for moral nudging.

DeGrazia believes that the overlapping consensus among competing views will be "fairly broad" (ibid.).[114] I remain agnostic on this. There will be many duties of beneficence, for instance, the value of which will undoubtedly be recognized, but that may fall into the category of principled non-enforceability. For reasons of caution, I will here include only two areas of moral duties that, I believe, more obviously fall into the category of enforceable and feasibility-constrained non-enforceable duties – behavior in traffic and environmental protection.

The grounds for singling out these two areas is the principle of harm – individuals have duties, often difficult ones to boot, not to cause or contribute to harm, and to prevent harm when this is not particularly costly to them. According to the WHO (2018), around 1.35 million people die from traffic accidents yearly, and another 20 to 50 million suffer non-fatal injuries, many of which end in physical impairment. And with regard to environmental protection, a study by Daniel Rothman (2017) suggests that human activities could contribute to the critical size of inorganic carbon in the oceans, causing mass extinction as early as 2100. Given the enormous capacities for

[114] Some authors are unconvinced by DeGrazia's assessment, or at least seem to believe the domain of principled non-enforceability is broader. Paulo and Bublitz assert that the consensus will be narrow because duty discharging with the help of external interference comes into serious conflict with fundamental rights: "Why is it legitimate to target the emotional dispositions of an outspoken racist? The right to hold opinions, including racist ones, without interference is among the most fundamental rights" (2019, 102). Constraints on direct interferences with beliefs do seem to be in order, as Chapter 6 will show. But Sparrow goes a step further and claims that the project of moral bioenhancement (and, by extension, moral nudging) presupposes a degree of elitism given that its leaders presume to "know what being more moral consists in". Any state that endorsed such a project "would thereby be committed to moral perfectionism" (2014, 29). But this would imply that the proclamation of any duty to be enforceable commits one to moral perfectionism. Surely, that is implausible. Persson and Savulescu correctly respond that "society could not function unless there was widespread agreement about moral norms to the effect that other citizens must not be killed, raped, or robbed of their property; that they should be helped when in need; that their good deeds should be reciprocated; and so on" (2015, 52).

harm in these two areas, they seem to be prime candidates for behavioral enhancement, which moral nudging can surely provide.

Now, I do not mean to suggest that consequentalist, deontological, and virtue-based views could not agree on other areas of enforceable duties. The fact that legal systems around the world share a propensity to protect similar kinds of rights and interests suggests a 'fairly broad' overlapping consensus. Other areas of duties that moral nudges can facilitate – anti-discrimination, organ donation, or equality in political influence – could find their way into the overlapping set. My choice is merely due to an intuition that failures to discharge duties in the two selected areas may produce consequences of such magnitude that any appeal to autonomy seems decisively defeated. Still, the reasoning about these duties does not reflect on the other candidates for the consensus. And even if there are conflicts of value in some areas in which we have duties, as I show in Chapter 6, it is not out of the question that moral nudges, albeit constrained by other values, may still be permissible.

Importantly, permissible nudges for enforceable and feasibility-constrained non-enforceable duties should not be wholly without institutional constraints. Two come to mind. First, there should be a constraint of proportionality. We should expect that the burdens imposed by moral nudges will be less weighty than what a proportionality calculation would allow. Turning once again to the debate on moral bioenhancement, Matthew Clayton and Andrés Moles argue that the permissibility of influences should depend on whether they only affect the beliefs and motivations about the relevant moral decision, and not others (2018, 247). Although no current moral nudge comes to mind that has this particular effect, we should certainly be cautious that the influences do not disproportionately disturb mental lives. Consider also Hazem Zohny's concern that the loss of autonomy an enhancement may cause could "entail a loss of an ability to lead a

prudentially good life" (2018, 272). Imagine, for instance, that a senior citizen with a meager pension starts giving away money to the point of destitution, under the influence of a green nudge. Moral nudge programs will have to make sure that beliefs and motivations are not burdened in these ways, either by abandoning nudges with such deep effects on mental lives, or avoiding nudge stacking. The difficulty of such control, however, is that it will often be difficult to establish that the changes in motivations were in fact due to a moral nudge.

Second, there should be a constraint of oversight. Allowing governments to engage in moral nudging covertly will undoubtedly raise worries that nudges will be misused for sinister and corrupt ends. Can we utilize covert behavioral influences while keeping an eye on government activities? In 3.6., the conception of watchfulness that is meant to defend personal autonomy included a societal condition – that a group of high-capacity experts are on the lookout for sinister scheming. A similar constraint on moral nudging is required here. Watchful experts must be able to blow the whistle whenever the government engages in dubious nudging. Some controversy will undoubtedly be raised about the dubiousness of the aim when such nudging is unveiled to the public, but this can be resolved by looking at whether the nudging reflects a deliberative and scientific consensus. In the case of protecting nudgees from self-regarding nudges, the whistleblower's job is not to make judgments about the good of the nudge, but to determine whether the nudge matches the nudgees' autonomous pursuits. Similarly, with moral nudges, whistleblowers observe duties that are backed by democratic and scientific consensus (when such ends exist), and raise alarms when nudges deviate from them. The presence of whistleblowers does not, and is not meant to protect personal autonomy here, but guarantees some accountability to an otherwise non-democratic part of the nudge project, and thus grants an opportunity for contestation.

Some might object that, to ground some enforceable duties in harm, I must not overlook the distinction between harm in general terms, and *wrongful* harm. Joel Feinberg argued that the first notion is non-normative, thus not all harms, as setbacks to welfare and interests, count as wrongs (1986b, 31-36). However, I will not adopt Feinberg's moralized view of harm here. It is true that harm which arises from wrongdoing could potentially outweigh the same magnitude of harm which does not, but it does not follow that the latter, at least in some cases, should not bear *any* weight.[115] Cases in which people's interests can be seriously harmed, through no fault or choice of anyone in particular, may sometimes at least give rise to enforceable or feasibility-constrained non-enforceable duties. Imagine that environmental degradation resulted exclusively from natural carbon cycles, and not in the least from human activity, but humans possessed means to alleviate the threat, if only they overcame collective action problems. Assume that Greg from a first-world country is not threatened, but Murray from the third world is, and Greg knows it. Greg seems to have an enforceable duty to help Murray by preventing the harm as best he can even though the harm does not result from wrongdoing.[116]

Let's turn briefly to principled non-enforceability. I treat cases of principled non-enforceability as those of *uncertainties* that are fundamentally moral. As Krister Bykvist states, "[w]e can be uncertain about how to weigh reasons of autonomy against reasons of benevolence, not because we lack an understanding of what these reasons are […] but because we are not sure

[115] There are, of course, other cases in which non-wrongful harm might not bear weight. If you were a better shoemaker than me, and your shoemaking skills and outputs put me out of business, this hardly seems a harm that requires interference.

[116] A serious consideration of behavioral influences will complicate matters for moral responsibility, in at least some cases. Given the vast number of cognitive obstacles that we are now becoming aware of, it will be difficult for a while longer to determine when humans are responsible for choices in which such obstacles get in the way. Such an assessment will have to wait for a more behaviorally-informed account of moral responsibility, which I leave for another occasion.

which of the reasons is more morally important" (2017, 1).[117] I spell out three different kinds of moral uncertainty, two of which block the permissibility of moral nudges, while the third remains open to them, but with significant constraints.

The first kind of uncertainty arises when we are not able to decisively derive a duty from clashing reasons for action. Each of the clashing reasons could decisively yield a duty if all other relevant moral considerations were absent. Otherwise, all we can gather is that the clashing moral reasons carry weight – that they are *pro tanto* reasons. Often, as in Bykvist's example, the weighing only delivers uncertainty about what our duties are. When duties are indeterminate, a cautious moral conclusion would be to suspend interference.[118]

The second kind of uncertainty occurs when a duty can be decisively determined, but we fail to determine the extent of protected choice granting individuals 'a right to do wrong'. Authors like Jeremy Waldron (1981) and Ori Herstein (2012) have claimed that a moral right to do wrong gives individuals a claim-right to commit decisively (and not merely *pro tanto*) wrong acts without being interfered with by others. Some areas seem to give rise to such a right. For instance, none of us should fail in our duties of promise-keeping or marital fidelity when there are no conflicting considerations, but it would be wrong to be interfered with committing these wrongs (unless, perhaps, we commit to being interfered with). In addition, it might be required that we muster our own motivation for the honoring of these duties to be worthwhile. In such cases, nudging seems impermissible. The uncertainty of this second kind is how far the right to do wrong extends, but

[117] A burgeoning literature has been dealing over the last two decades with the problem of moral uncertainty and resolving deadlocks between moral theories. See, for instance, Lockhart (2000), Sepielli (2013), and Bykvist (2017).
[118] We could still interfere against really bad options. Imagine that there was an uncertainty between A and B, while C was not only inferior to both, but judged to be non-controversially bad. We could then interfere against C. If there were multiple interfering options against C, we would do well to pick the one that does not favor A or B as a side-effect. I thank Andrés Moles for pointing this out to me.

even the advocates seem to recognize its limits. Herstein states that we are not afforded with a "pervasive right to do any and all wrong […] particularly egregious wrongs" (ibid., 347).

Finally, the third kind of uncertainty concerns *how much* some duty requires once established at least minimally, given conflicting considerations. Persson and Savulescu seem to refer to such cases when they say:

> "although in some situations we do not know *exactly* what morality requires of us, we know that it requires *more* than what most of us do, for example, that it requires to give more aid to the needy in developing countries than most of us actually do. Strengthening these dispositions would contribute to this end." (2015, 52-53; emphases in original)

This will be the main concern of Chapter 6, where I claim that the third kind of cases of uncertainty, an example of which is charitable giving, should permit the facilitating of duty discharging via moral nudges, but that these ought to be constrained so that conflicting values, mainly personal autonomy, can be protected.

5.3. The *pro tanto* wrongness of moral nudges – considerations from the moral bioenhancement debate

I will maintain throughout the remainder of the chapter that the benefits of moral nudges are sufficient to outweigh the regrettable side effects that they might cause. Yet, a nonpartisan evaluation before making the all-things-considered judgment requires these other considerations be given closer attention. Like before, much inspiration about the normative assessment of moral nudges can be taken from the moral bioenhancement debate.

For starters, the two debates standardly endorse similar ways for conceptualizing influences. As I mentioned in the Introduction, Thaler and Sunstein use a broad conception of nudging that admits any tweaks to the choice environment regardless of whether they trigger a heuristic or tamper with motivations, making reflective influences such as the GPS or the 'look right' road sign in the UK to be included. The narrow conception, on the other hand, only observes non-reflective influences, such as heuristic triggers. Interestingly, the reflective and non-reflective influences observed under the two different conceptions of nudging closely match the categories of moral bioenhancement. According to Schaefer, for instance, an indirect moral enhancement "is designed to make people more reliably produce the morally correct ideas, motives and/or actions," while a direct moral enhancement "is designed to bring someone's beliefs, motives, and/or actions in line with what the enhancer believes are the correct moral beliefs, motives and/or actions" (2015, 262). The first kind of bioenhancement looks to enhance the dispositions relevant for arriving at sound moral judgment (which brings about moral compliance rather than mere conformity), whereas the other bypasses those disposition, or merely boosts the motivations that do not improve moral judgment itself.[119]

Note that my aim is not to compare all possibly relevant properties of moral bioenhancement and moral nudging, although such properties might affect which influences we find more obviously permissible or impermissible, or which would be preferable when choosing a policy. I leave this for future research. The aim here is only to highlight familiar objections against

[119] A relevant difference between indirect moral enhancement and reflective influences under a broad conception of nudging might be that while the former improves the moral agent's dispositions to make judgments, the latter improves the choice context to make reasons more salient in order to better interact with the moral agent's standard dispositions. While endorsers of moral enhancement would likely find reasons to endorse such reflective influences from their own argumentative accounts, reflective influences do not 'improve' agents in the strict sense that indirect enhancements do.

moral bioenhancement, and then if the two techniques share the properties targeted by those objections, apply them in assessing moral nudges.

Although the endorsement of moral bioenhancement is more commonly grounded in indirect enhancement, it is the direct enhancements that standardly come under critical fire. Assuming that heuristics-triggering nudges and direct enhancements are similar in morally relevant terms, I once again focus on narrowly conceived moral nudging and scan the objections from the moral bioenhancement debate that could be relevant for its permissibility. For instance, an oft-cited objection against direct bioenhancement is that, by tweaking attitudes so that some choices are considerably more probable than others, it denies individuals a 'freedom to fall' (Harris 2011). For now, let's take this objection only to refer to the capacity of agents to do otherwise, in the face of some actions being made more probable. The very purpose of nudges is to make some actions more probable than others. Let's also assume, for now, that nudges make certain actions *only somewhat* more probable, without effectively closing off alternatives. In such circumstances, nudges would make nudgees no less free than the plethora of psychological influences that divert people from acting on a duty of environmental protection. If nudges making nudgees somewhat more likely to do A rather than B is sufficient to curtail 'freedom to fall' (doing B), then pervasive psychological effects making individuals somewhat more likely to do B rather than A would suffice to curtail 'freedom to succeed' (doing A). The fact alone that some options are more likely to be chosen than others by virtue of non-reflective influences, without options being closed off, should not count out nudges, given that most choices will be characterized in such a way. Consider Persson and Savulescu's point that if morally bioenhancing one's altruism makes one unfree to fall, then women, who have a greater capacity for altruism than men, would be less free (2012, 111-112). Such a conclusion, as they say, is obviously implausible.

Instead, moral nudging could be conceived as making actions significantly more probable, effectively or at least nearly closing off alternatives. If so, moral nudges bear some resemblance to the 'God Machine' (GM). Persson and Savulescu's GM is conceived as a super-computer which monitors people's thoughts, beliefs, desires, and intentions, and is able to modify these without their noticing. The modifying occurs when people decide to commit egregious wrongs, like murder, rape, and torture. If such decisions are made, the GM simply initiates a change of mind, which individuals would not perceive as alien to them. There would be no intervention into less harmful acts, like dishonesty or promise-breaking. People would enjoy full non-interference as long as their decisions did not lead to a 'fall' of the worst possible kind (Savulescu and Persson 2012, 412-413).[120]

Given its enormous benefits at the cost of undercutting freedom in a very limited range of cases, some authors believe that the GM easily outweighs considerations of freedom (for example, DeGrazia 2012, 367; Savulescu and Persson 2012, 414; Persson and Savulescu 2016b, 275, 277).[121] In Pugh's words, objections against the GM, as well as direct enhancements of lesser impact, "must render plausible the idea that allowing immoral actions is a price worth paying for preserving freedom" (2017, 74). Similar to my case for moral nudges, Persson and Savulescu argue that individuals should bind themselves to the GM, much like we bind ourselves to coercive laws which prevent egregious harms (2016b, 276). There are, of course, possibly relevant moral differences between moral nudges and the GM. Moral nudges would presumably neither be as reliable nor as effective, they would interfere "less directly" with the person's mental states, they

[120] The idea for this kind of intervention is borrowed from the work of Frankfurt (1969).

[121] Oddly enough, Persson and Savulescu always make sure to point out that the GM does not count as a moral enhancement (2012, 414; 2016a, 265; 2016b, 275), and that modifications of motivational dispositions only count as moral bioenhancements in a wider sense (2019, 11). Yet, as I mentioned, the main front between endorsers of moral bioenhancement and their critics is on direct enhancements, which the former standardly defend as an extension of their position. It should not matter much, then, that moral nudges are not "moral enhancement proper".

would not purge all trace of contrary conviction, and an oversight constraint against them is much more easily conceived than against the GM. Still, a freedom-based objection against the GM seems fairly translatable to the context of assessing moral nudges. I now turn to two other inherited and more weighty objections, grounded in moral compliance and value uniformity.[122]

5.3.1. ***Moral compliance***

Harris's concern does not seem to be solely about freedom in the sense of options remaining open, but about the undermining of our engagement with moral reasons for action. Moral bioenhancements (and, I will presume here, moral nudges), according to Harris, make doing "immoral things impossible", and stand in opposition to giving people "moral, legal and prudential reasons to refrain" (2011, 105), while, for Michael Hauskeller, they "turn us into mere puppets hanging from strings" (2013, 53). The ability to act on moral reasons seems to be a constitutive part of the Rawlsian moral power of having the capacity to understand, apply, and act from the public conception of justice.

An important distinction is yet to be drawn here between *conforming* and *complying* with moral reasons for action. To conform with a reason, we need only act in the direction that the moral reason would favor, whereas to comply requires our action to be guided specifically by the moral reason, or in accordance with *why* the action is favored by the reason (Raz 1999, 178-179). Imagine, for instance, that Vivian is dreaded by the mere sight of doctors and can barely get through the anxiety of a standard medical check-up. Suppose that Phil, her friend, can allay her anxiety by accompanying her to her general practitioner. Phil's company would be characterized

[122] We could, of course, raise important objections to moral nudging that are not inherited from the moral bioenhancement debate. For instance, moral nudges might be unfair since they could produce inegalitarian burdens. Because certain individuals often have strong preferences that allow them to fend off nudges, or are particularly vigilant to non-reflective influences, they may take up less burden in the discharging of collective duties. I tackle this objection on the example of charity nudges in 6.3.1.

as compliance with a moral reason if he was driven by the fact that his action could help Vivian in distress. It is only conformity with a moral reason if he is, for instance, driven by the opportunity to use the situation as an excuse to drop off his children at his mother's and spend an afternoon in the waiting room reading a good book.

Does moral nudging undermine compliance with moral reasons for action? Not necessarily. Some instances of motivation boosting could plausibly highlight relevant considerations for moral compliance, like Oxfam-style pictures of destitute children that move us to donate. Such interventions do not seem particularly threatening for considerations of compliance, unless individuals want to bring themselves to become compliant without being deliberately influenced by others. I will put these interventions to one side here. On the other hand, we imagine that bypassing reasoning to behave in a particular way hardly fosters moral compliance. This subsection discusses the correctness of such assumptions.

Let's look first at whether a moral nudge undermines the engagement with moral reasons for dissenters – persons who disagree with the direction of the nudge. Suppose that a person makes some moral judgment, for instance, that it is justifiable to be a reckless driver. Under the influence of a number of traffic nudges, the person behaves far less recklessly, unbeknownst to him that this was due to non-reflective external influence. Certainly, the person acts non-compliantly under the influence of the nudge. We might think that governments should not interfere with people's judgments about whether they should abide by a moral rule, and that it would be valuable to leave the dissenter to his own moral reasoning. But neither of these points are morally obvious. In legal practice, for instance, the sentence that a person receives for a serious offence may often depend on whether he expresses regret for not abiding by moral reasons for action. Moral nudging here

would undermine engagement with reasons for actions, although the reasons are of a kind that are not obviously worth protecting.

Now let's assume that a person holds a moral belief that he should, for instance, contribute more to environmental protection or poverty alleviation, and is willing to bind himself to some influence that would jumpstart his motivation. Imagine that this person would not be able to muster the motivation to act on his belief, if he is not nudged. Would the moral nudge undermine his engagement with moral reasons? This depends on what we take to be required for moral compliance. We could think that compliance requires that moral reasons "vividly resonate" with the agent at all times whilst performing the action or a set of actions. But this seems too demanding. Consider the following example by Douglas:

> "Suppose I believe that I ought to be more moved by the plight of the global poor, and ought to do more to help them. However, I have trouble drumming up much sympathy for them. To remedy this, I set up my television so that it regularly displays disturbing and graphic images of the effects of poverty, though for such brief periods that I do not consciously recognize them. Nevertheless, through subliminal effects, the images increase my feelings of sympathy" (Douglas 2013, 163).

The judgment that I 'ought to be more moved' (although I am not) and binding myself to action by setting up my television seems sufficient for subsequent actions to be qualified as arising from moral compliance, i.e., for me to be guided by the moral reasons. If so, then binding oneself to moral nudges would not represent a loss in terms of engaging with moral reasons.[123] Sustaining the aforementioned Rawlsian moral power could, thus, take on a less direct form. Of course, objections could be raised. It could be claimed that this sort of 'distant compliance', which does not commit agents to any subsequent engagement with reasons for action, is not sufficient to

[123] See also Savulescu and Persson (2012, 414), who claim that binding to the GM would come at no loss to autonomy.

safeguard morally unified agency in the long run. Agents need to consciously engage with reasons for action in order to constitute their moral identity over time. I return to this objection shortly.

Finally, what about persons who believe they ought to do more for the environment, but fail to do so, and do not commit explicitly to being morally nudged? Does nudging undermine their compliance? My take is that it leaves their (lack of) compliance *no worse* than it would have otherwise been. The motivation of many of these individuals was already being undercut by detrimental influences (such as planning fallacy, hindsight bias, or scope neglect), hindering them to act compliantly in the first place.[124] If so, then a failure to engage with moral reasons in the subverted choice context is a mere continuation of a previous failure, but one that at least secures moral conformity. Would this case qualify as distant compliance? This is difficult to assess. Different degrees of motivation will be exhibited among members of the group. It would be false to call a barely motivated person compliant, or a considerably motivated person only conforming, which is further complicated by the fact that persons are motivated by factors other than their moral beliefs.

Hence, moral nudges introduce a clear non-compliance scenario in the case of dissenters. In the other two cases, they ensure conformity, and, it could be argued, a distant compliance. But a different objection might suggest that the presence of nudges undercuts agents' opportunities for 'vivid' compliance. Now, it could be the case, especially with those who bind themselves to moral nudges, such as Douglas's philanthropist, that they will stop bothering to actively engage with moral reasons, knowing that they can fall back on a motivational safety net. But as I mentioned

[124] It could be claimed by some that the nudge undermines their opportunity to build up motivation by themselves. Some individuals might even prefer to fail at their duty than being helped to conformity. They might believe that moral acts are worthwhile only if they are carried out independently from external interference. I talk more about this with reference to charity nudges in 6.2.1.

earlier in the chapter, it would be wrong to conceive of moral nudges as purging all moral reasoning or brainwashing people into alien beliefs. People who acknowledge their enforceable and non-enforceable duties and are able to garner the motivation to act on them do not act less compliantly because of the presence of moral nudges.

One final objection might hold that distant compliance might be preferable to disregarding duties, but that this is insufficient to realize a morally unified agency (one that, perhaps, sustaining the Rawlsian moral power requires). Christine Korsgaard (2009) has used the concept of self-constitution to characterize a morally unified person, who is required to form and continuously act from principled ends to sustain moral identity and personhood. Following Korsgaard, Steve Matthews asks: "does the enhancement promote in the agent a capacity for responding to reasons that enable morally unified agency, or does it disrupt this capacity?" (2015, 109). Assessing environmental nudges in view of Korsgaard's conception, Schubert worries that nudges are not compatible with an agent's self-constitution, given their tendency to discourage active choice (2016, 25). So, if active choices are required for agents acting on their principled ends, and if choices prompted by nudges are not in the proper sense 'active', then morally unified agency requires vivid, and not just distant compliance.

Unlike Schubert, I am not convinced that Korsgaard's account requires vivid compliance across the board as a condition for unified moral agency, lest it becomes overly demanding and somewhat psychologically naïve. Allowing some moral choices to be arranged between agents and choice architects for the purposes of making behavior more effective in discharging moral duties hardly seems sufficient to constitute a breach in the agents' morally unified agency. Otherwise, people would likely suffer breaches if they could not live up to their principled ends solely due to their psychological shortcomings. Remember also that I argue for moral nudges only for

facilitating duties that are sure to be found in an overlapping consensus. Given that this is a limited range of cases, it seems unlikely that these nudges would breach a morally unified agency, or (as the previous objection fears) eliminate all opportunities for vivid compliance. So, moral nudges do hold the capacity to undermine moral compliance, but these consequences are neither certain nor necessarily grave.

5.3.2. ***Value uniformity***

Schaefer raises a Millian concern[125] about direct moral enhancement possibly threatening moral disagreement, a freedom to hold dissenting opinions and criticize others publicly (2015, 263-264). He believes the soundness of Mill's arguments speaks "powerfully against widespread direct moral enhancement" (ibid., 264). The value of moral disagreement is that it keeps democratic dissent alive, which is the only way in society to come to revise established beliefs that are false, and maintain the pursuit for moral progress (Mill 1859). The effect of direct moral enhancement, Schaefer thinks, is that the inculcating of moral ideas and motives that it consists in will undoubtedly rule out dissenting opinion, tending significantly towards value uniformity with the enhancers (2015, 266).[126]

Would Schaefer's objection to moral bioenhancement extend to the permissibility of moral nudging? The knee-jerk intuition might be that it would not, at least given how Schaefer conceives

[125] He draws his inspiration for the argument from John Stuart Mill's 'On Liberty' (1859), particularly Chapters 2 and 3.

[126] Schaefer's worry is possibly reinforced by Persson and Savulescu arguing that "[n]on-democratic, authoritarian forms of government are better placed than democracies to implement unpopular reforms effectively" (2012, 86). They note that China was, compared to India, much more successful at implementing demographic reforms such as the one-child policy due to its non-democratic institutions (ibid., 87). Hence, Persson and Savulescu believe China is also much better placed to secure "effective implementation of environmental-friendly policies" (ibid., 88) than Western countries. This may give the impression that moral bioenhancement is only tenable and should be carried out in strictly non-democratic conditions.
Some may think, contrary to Schaefer, that the suggested potential of moral nudges to discourage dissenting opinion actually speaks in favor of such nudging. See Engelen et al. (2018) for the capacity of nudges to facilitate moral learning, and Bovens (2009, 215, 217) for habituation of nudged behavior.

of moral bioenhancement. As I maintained, it seems wrong to suggest that bypassing reasoning and boosting motivation produces an effect on beliefs that would be accurately characterized as "inculcating ideas" (ibid.). The worry about value uniformity might be derived specifically from the suggestion that direct enhancements inculcate ideas, while blocking dissenting ideas from being adopted.

While empirical work is yet to be done on how exactly evolving biochemical and behavioral technologies stimulate conforming action, and whether their effects on mental lives are comparable, I have maintained that moral nudges would not have a direct 'inculcating' effect. This is not to deny that moral nudges *could* produce value uniformity by, for instance, contributing to the forming of strong social norms and taboos. If they did, such outcomes might be regrettable. They could negatively affect the democratic and possibly scientific deliberation required to inform moral nudging. And if so, they could block the full realization of the moral power to understand the public conception of justice. But nudges should not be conceived as blocking dissent and general deliberation about social aims altogether. In fact, in Chapter 3, I have claimed that a general deliberation about aims – the democratic condition – is what characterizes a watchful society in which ex post transparent and non-transparent nudges are permissible.

I will make five further points about why the worries about value uniformity and a democratic deficit might be overblown. The first two points respond to Schaefer's claims directly, while the last two points refer specifically to how moral nudges fare against the objection. First, and most importantly, while moral nudging and direct enhancement alike could produce regrettable effects, it is not clear why the badness of uniformity in certain areas would come close to the badness of ultimate harm. Schaefer responds to this line by drawing from Mill, who stated that presumptions of certainty (about what counts as very bad) are all too common in history, and

that even the enlightened Roman emperor Marcus Aurelius thought doctrinal unity justified the brutal suppression of Christianity (1859, 30-32). But surely Schaefer would accept that, with sufficient evidence, the urgency of a problem and its capacity for harm could at some point outweigh the requirement for maintaining moral disagreement. My discussion here is about why moral nudges could be regrettable, and why moral disagreement could count against nudges in conditions of moral uncertainty, but I do not think the concerns altogether undermine their permissibility for facilitating at least some enforceable duties.

Second, while I agree with Schaefer that direct enhancements, and hence, moral nudges, could have adverse effects on moral progress, it is not clear why moral progress should consist *entirely* in preserving moral disagreement. This might be due to what Schaefer sees as the function of moral disagreement and his optimism about moral progress: "If we expect moral ideas held by the public to, by and large, become more and more in line with the truth, then there is very strong reason to want to preserve the ability of moral ideas to evolve" (2015, 265). Let's grant for the sake of argument that disagreements linearly tend towards discovering moral truths. It would still be perfectly conceivable for societies and their members, to be proficient at *finding out* moral truths, but to do poorly in *acting on* what such truths require. A common understanding of moral progress in society, I believe, also concerns overcoming the moral wrongs of the past, rather than only elucidating what the content of those wrongs is. Schaefer himself recognizes that it is possible to have too much moral deliberation and that we should know when to call them to a halt (ibid., 285). However, this knowledge would warrant moral progress only if we endorse a strong motivational internalism – that these correct moral beliefs will garner sufficient motivation to bring about action. But if strong motivational internalism is false, as I have suggested, then moral disagreement is insufficient for acting on moral truths, and thus for moral progress. What might

be required are direct bioenhancements, or direct behavioral influences. This is not to suggest that we need to choose between moral disagreement and direct influence. The cautious account that I have defended advocates moral nudging only for facilitating established duties, while steering clear from areas of moral uncertainty. Even if nudges conceivably stifle dissent, it should not be assumed that moral disagreement will be undermined across the board.

Third, I have so far granted that moral nudging produces value uniformity similarly to Schaefer's conception of direct moral enhancement. But this is a mere stipulation. We should recall from Chapter 2 that nudges are based on exactly the same brute effects that non-designed influences have on our cognitive heuristics. But if this is the case, how is it that these widespread non-reflective effects which are *not designed*, predictable enough as 5.1.1. shows, do not produce value uniformity? Why is it that we hold vastly different views about problems of climate change and global poverty? And if so, why should we expect value uniformity to be an effect of moral nudges?

Fourth, I have presented moral nudges as a non-democratic policy tool. But the non-democratic aspect relates to their covertness and non-transparency. Unlike self-regarding nudges, which I have labelled at times as 'democratic', moral nudges are not curtailed by the constraints, like transparency, that earn them the same attribute. Still, the fact that moral nudges are non-democratic does not rule out democratic deliberation about the values that they promote. This leads me to my fifth and final point – moral nudges are not the be-all and end-all policy. If they undermine moral disagreement about important values, then governments can promote democratic deliberation via other policy means. For instance, John et al. believe people should be encouraged into "participatory budgeting, the mini-publics of citizens' assemblies and juries, and online

forums enabled by developments in information and communications technology" (2009, 365), as part of a broad deliberation strategy to reflect on domestic and global moral dilemmas.

5.4. Strengthening the case for moral nudging

5.4.1. *The inevitability argument*

In 5.3.1., I have confirmed that many moral nudges, unless individuals commit to them, fail to improve the responsiveness with moral reasons for action, although they might not make it *any worse either*. In these latter cases, moral nudges merely replace influences that were working as a psychological detriment for agents to meeting enforceable and non-enforceable duties. But it could be objected that I have failed to note that the latter kind of influence is more morally threatening, given that moral nudges are intentional, whereas the influences they replace are not. This is a familiar concern, one that I have detailed at length in in 2.4.2. and 2.4.3., and it is from there that I will draw a familiar solution.

We should once again recruit the talents of Carolyn, the expert choice architect. Imagine that Carolyn could employ a moral nudge that would get John to invest money or time into an environmental cause. If John were to invest, however, he would do so non-compliantly, or with distant compliance (depending on the reader's theoretical preference). Let's assume also that John is by no means opposed to the environmental cause. He sincerely responds in the affirmative if asked whether he thinks he should do more. Yet, his conviction has so far fallen short in mustering moral motivation that would get him to act. Reasoning with John has not helped either – he affirms many of the reasons presented to him and agrees that they are sufficient for action, but changes in his discharging of moral duties have been next to non-existent. To top it all, his affirmation of

moral reasons has so far not led him to bind himself to any of the aforementioned motivational mechanisms. As some empirical data in 5.4.3. will show, John is the *everyman* – the description of his moral beliefs and motivation (or a lack thereof) is one in which many of us can recognize ourselves.

If Carolyn were to employ the moral nudge, she would, I claimed in 5.3.1., certainly not improve John's moral compliance (at least not in the vivid sense), yet she would not worsen it either. John would have acted non-compliantly due to the shortcomings of his psychology or an unsuitable choice environment, given that, *ex hypothesi*, facing him with reasons would not make a difference. Still, the complaint is that Carolyn introduces intended influences where there were previously none. Lest we forget, however, Carolyn is a champion among choice architects. She has a near-perfect grasp of all the cognitive failings described in 5.1.1., and many others. It is predictable to her what will happen if she does not intervene.

In Chapter 2, I have claimed that if the effects of possible environments are fully predictable to a choice architect like Carolyn, it would be difficult to see what, if anything, would explain the moral difference between her intervention and her omission to intervene. I claimed, and maintain, that acts and omissions are morally 'close' for expert choice architects. With regard to moral nudging for environmental protection, the case against the acts/omissions doctrine is particularly strong, given that there is nothing science-fictional about the insights choice architects can have about psychological causes that explain failure of action in the area of environmental protection. And if choice architects can predict the behavior of those affected, together with the harm that is implied, then they bear responsibility for omissions the same as they do for active interventions. The closeness between acts and omissions dispels the worry that there is something more troubling about designed, as opposed to non-designed influences. New insights about

cognition and heuristics are akin to technological breakthroughs, and these bring about new responsibilities, namely of failing to prevent. As Persson and Savulescu state:

> "the more extensive the powers of action that we possess thanks to scientific technology, the greater our moral responsibility. More precisely, if we are aware that it is in our power to prevent some harm, and we refrain from so doing, we could be as responsible for its occurrence as we would be had we knowingly caused it." (2012, 62)

5.4.2. ***Scarce resources***

The resource-based account of autonomy that I have defended in Chapter 2 is premised on limited reflective capacities. At the core of the resource-based account is that individuals organize their reflective capacities in ways that suit their management styles, but within constraints of psychological feasibility. I claimed in Chapter 3 that self-regarding *ex post* transparent and non-transparent nudges can relieve our cognitive exertion when we agree with their direction, and thus enable us to use limited reflective resources on projects more meaningful to us.

With personal autonomy, the range of available management styles is wide – I have not mentioned many constraints on balancing out reflective and non-reflective capacities. But if discharging some moral duties requires compliance in the vivid sense, then that entails investment of at least some reflective resources, as a constitutive part of discharging those duties. To keep a promise to a friend or tend to him when he is distressed will require our full attention. Or, we might think that cases of moral uncertainty require that we guide our actions in response to properly engaging moral reasons. Morality will thus act as a significant additional constraint on how we organize our reflective and non-reflective capacities, given the wide range of duties that we owe others.

However, as I claimed in Chapter 2, the fact that we want to carry out some activities at low capacity does not necessitate that we are indifferent to them. Rather, we might feel so certain

about the values underpinning some actions that we doubt that performing them reflectively would lead to value revisions. Or, we might believe that executing actions at low capacity might improve our efficiency.[127] Similar arguments can be made about engaging with some moral duties. If such duties are open-and-shut cases, and agents could discharge them much more effectively at low capacity than upon engaging with moral reasons, then agents might be morally required to do so.

Another consideration of limited reflective capacities could favor moral nudges. I have suggested earlier that moral nudges do not improve moral compliance, but in many cases, they do not worsen it. Yet, moral nudges could improve moral compliance in one sense. If agents have limited reflective capacities, which I have claimed throughout the dissertation, and moral compliance requires investing reflective capacities, then agents have limited capacities for moral compliance. If some duties are open-and-shut cases, then we would do better to invest reflective capacities into moral cases which are not. Hence, nudging open-and-shut cases would presumably help saving up the limited resources for cases that require moral reflection.

Many, if not most of our moral acts, or omissions to act immorally, fit this particular bill. Although there are important moral reasons with which we could comply, we usually find it sufficient to merely conform. Joseph Raz argues that although we permanently have moral reasons not to kill those around us, we rarely ever engage with those reasons – the thought simply never occurs to us (Raz 1999, 181). Raz believes that, with the exception of special circumstances, moral acts do not lose their worth if they arise from conformity; it is most important that acts for which there are available moral reasons get done (ibid., 180-182). And if moral compliance is saved up for special circumstances, then it should be permissible that agents should perform many of their

[127] For this point, see Christman's argument from 2.3.1.

moral duties at low capacity, and allow them to be facilitated by moral nudges. This should also imply that the aforementioned moral power will require reflective understanding of the public conception of justice, but should allow, if there are no special circumstances, that we apply and act from it without vivid compliance. Whether circumstances are special should be, in part, determined by considerations of reflective exhaustion.

Lastly, if we care deeply about allowing individuals to conserve their reflective resources, then we should take the previous argument one step further. We should think that sparing individuals from laborious moral reflection goes not only to the point of enabling them to use their reflective resources on controversial moral cases, but on their own personal projects. Requiring people to use up all of their conscious mental life on moral reflection would be to demand moral sainthood. Thus, the argument for moral nudges not only ensures the investment of reflection on moral cases that matter, but should also save individuals from moral preoccupation and allow them to realize their personal projects.

5.4.3. ***Personal autonomy***

Finally, people will often endorse moral goals as part of their conceptions of the good life, but might find it difficult to pursue them. They might have trouble overcoming psychological barriers even when these are pointed out to them or act on reasons even when they re-engage with them. Could moral nudges promote personal autonomy, in the absence of conditions of watchfulness?

On my account of resource-based autonomy, they could. Individuals could endorse the pursuit of some moral goal, say, environmental protection, by developing certain habits or accepting choice environments in which they could attain their goals while operating at low

capacity. Or, they might desire investing reflective capacities to pursue the moral goals, but upon seeing that their pursuit has failed, they might still prefer being nudged into conformity than not accomplishing the goal at all. An elaborate analysis of how one version of moral nudges, namely, charity nudges, fares against considerations of personal autonomy on a resource-based account will be provided in the next chapter.

For now, I will offer some good empirical bullet points, giving reasons to suspect that the numbers of those failing to act on their goals are considerable in liberal democracies. Daniel Pichert and Konstantinos Katsikopoulos (2008), for example, prove that most citizens lend support to environmental policies like green energy, but fail to act on their testimonies due to the presentation of information and decisions that they face. Looking at the stated willingness of respondents to pay for public goods, such as preserving bird species or preventing oil spillages, Kahneman et al. (1993) note that in hypothetical markets people commit to impressively large amounts. And citing a study from Iowa, Thaler and Sunstein note that people's stated willingness to become donors often fails to translate into action – of 64% who stated their willingness, only 36% signed an organ donor card under an explicit consent regime (2008, 176).

Intuitively at least, one of the more mysterious cases of moral action falling through is charity donation, given that, once again, most people agree that the Western world should do much more to alleviate world hunger and global poverty. I will treat charity donation as an area of permissible nudging, but one in which global duties need to be balanced out with considerations of personal autonomy. I turn to spelling out an account of how such value balancing might work.

Chapter 6: Charity Nudges

In 2013, the Behavioural Insights Team published a report titled 'Applying behavioral insights to charitable giving'. In it, the BIT demonstrates the workings of behavioral techniques with notable potential for encouraging people to give to charity. The growing compilation of separate tests and trials also seem to support this finding. In the context of our hopes of ousting, or at least alleviating global poverty in the third world, these are very exciting news, but they come with a caveat. It consists of the threat that behavioral techniques may represent for the individual's autonomous choice. This chapter discusses the special case of charity giving as that of *value uncertainty*, a notion that I have introduced in the previous chapter. I mentioned there that three different kinds of value uncertainty exist. Charity giving, I argue is the third kind – what is uncertain is the extent of the enforceable duty's requirement once it has been established at least minimally and how it weighs against conflicting considerations, like those of personal autonomy. Whether moral nudges that facilitate charity giving will be permissible is determined by whether they can accommodate autonomy considerations. Additionally, even if we could determine the exact extent to which the duty of alleviating global poverty is enforceable, our circumstances are non-ideal in two senses – people fail to live up to their obligations, and institutions are ineffective. This is the normative background with which charity nudges will be assessed.

The collective moral duty of alleviating global poverty is endorsed by most contributors in the philosophical literature (Brock 2009; Singer 2004; Pogge 2007), though its character remains elusive. My contribution in this chapter will be to suggest that nudging, carried out by charity organizations and sometimes in tandem with governments, may be used as a vehicle for meeting this elusive duty in the face of cooperation problems and information deficits. I argue that nudges are called for in cases where coercive measures have done too little on both state and international

levels for the alleviation of poverty even with consensus on the existence of normative duties, if not on what exactly these duties require us to do. I will claim that individuals retain the right to follow their conceptions of the good and pursue them accordingly, but in view of their share in the collective duty, they are without a right not to be influenced by nudge techniques. The social endorsement of such a view will give rise to a *nudge ethos* with regard to charity donations.

First, I give an overview of nudges that have already been tested for these purposes, and mention the cognitive heuristics that are usually targeted. Secondly, I discuss Meena Krishnamurthy's claim that utilizing nudge techniques most likely includes a violation of personal autonomy of at least certain individuals (2015), and discuss how much her argument manages to prove, given the biases diverting us from discharging our duties for poverty alleviation. I then move to the third part, where I outline the conception of a nudge ethos, and present two of its versions – a lax and a strict one. Finally, I face both versions with some noteworthy objections.

6.1. The potential of charity nudges

In this section, I demonstrate the capacity of nudging for affecting behavior in the area of charity donations. I will call these techniques *charity nudges*, and will consider them a subtype of moral nudges. While examining the different studies on the effects of charity nudges, I also make note of the cognitive heuristics which are at work behind their efficacy. As mentioned continuously throughout Chapter 5, relevant heuristics may often take effect spontaneously, without any nudges used. In that case, the effects of heuristics may often deter individuals from donating even when they are determined to do so. So, once again, the relevant dilemma will not be whether to trigger people's heuristics with nudges or leave them to their reflective deliberation. Rather, it will be

whether to leave people to the effects of certain cognitive heuristics that are triggered with no interferences from others whatsoever, or to trigger heuristics to produce morally desirable outcomes. The heuristics that I mention are not exhaustive, but are the most relevant for donation contexts, and are meant to tease a bigger picture of actual and possible behavioral effects in the area of charity donations.

6.1.1. ***Studies and heuristics***

BIT's study report (2013) consisted of five trials.[128] The first divided the donors of the Zurich Community Trust (ZCT) into three groups, the members of which were asked in different ways whether they would want to increase their charitable giving in the following year. While the members of the first group were asked for a one-off increase over the course of a year, the members of the second and third group were asked to increase giving at the same rate after the one-year period. Furthermore, the difference between the second and third group was in different suggested amounts (£1/2/3/5/10 vs. £2/4/6/8/10). The study showed that the members of the third group donated a much higher amount than those in the other two groups.

Anchoring explains why the group with anchors set higher (£2/4/6/8/10) donates a higher net amount than the group with anchors set lower (£1/2/3/5/10). In another study, Karlan and List (2007) tested whether indicating the percentage of some designated donation goal to the potential donors affects how much they will end up donating (say, whether people will be more easily drawn to the donation scheme and whether they will donate more money the closer the percentage is to the designated goal – for example, at 75/100, rather than at 10/100). The study shows people are

[128] Not all of these studies are concerned with donating for the alleviation of *global* poverty, which will be my primary concern in the chapter. It is evident, however, that the nudges portrayed can easily be re-appropriated with this aim in mind.

not only more likely to enroll into the scheme if the percentage is closer to the target, but they also give much more money.

The second trial tested how an opt-out technique affects enrollment into donation schemes. Small changes were made on the payroll giving forms for the employees of the Home Retail Group, having as an effect that employees were enrolled into the donation scheme by default unless they stated otherwise. The number of donors rose from 6% to 49%. In other words, if employees are enrolled into a donation scheme unless they state otherwise, with full knowledge of both the automatic enrollment and the opt-out option, they are far more likely to stay enrolled in the donation program than if the default option is not to donate. As indicated many times throughout the dissertation, such trials are clear showcases for the status quo bias.

In the third trial, employees of HMRC were sent two different kinds of cards from other HMRC employees who donated to charity. While all cards consisted of a message from the employees explaining why they give to charity, half of them contained a picture of the person writing. The trial reports that 6.4% in the picture group enrolled, more than twice as many as in the control group (2.9%). The results of this trial are explained by the so-called peer effect, which makes individuals more likely to donate if they see people like themselves donating as well.

The fourth trial tested whether employees of Deutsche Bank in London were willing to donate a day's salary within the span of a year. They were sent an e-mail from their CEO, and different combinations of techniques were tested – some were addressed in the e-mail by their first name, while some were greeted by volunteers and given flyers and candy when they came to work. The study shows that those who were addressed by their first name in e-mails and greeted by volunteers enrolled in highest numbers (17%), while those exposed to neither stimulus enrolled in

lowest numbers (5%). This trial combined the knowledge of two behavioral insights: people's tendency to respond positively when addressed in a personalized manner, and their strong tendency to reciprocate when given a gift.

Finally, the fifth trial tested different techniques of prompting people to leave money in the process of arranging their wills. In the control group, which was not prompted at all, 4.9% gave to charity. Of those who were merely asked whether they would like to give, 10.8% donated. Lastly, those who were told that many others leave donations enrolled in highest numbers and left the highest amounts (15.4%; they gave twice as much as the other two groups). This is a great example of, on one hand, utilizing the status quo bias which usually makes people too inert to include donations into their will, and on the other hand, taking advantage of social norms.

Other studies similarly point to the impressive potential of charity nudges for encouraging donation. Research by Francesca Gino and Sreedhari Desai (2012) shows that individuals are much more likely to donate if their childhood memories are evoked prior to being prompted to donate. This is an example of priming – setting up the cognitive and emotional context in which individuals make later decisions. The authors argue that evoking childhood memories stimulates feelings of moral purity which then, in turn, may prompt a variety of pro-social behaviors.[129] Finally, an experiment by Eileen Chou and J.K. Murnighan (2013) demonstrated that carefully framing a message for attracting blood donations makes a notable difference. Their results suggest that the percentage of donors rose more substantially (1.31%) when the call for donations was framed in terms of loss ('prevent a death') rather than in terms of gain ('save a life' – 0.78%). Such uses of

[129] Aside from encouraging charitable giving, the authors also show that priming encourages feelings of empathy for those in need, helping others with simple tasks, and judging morally questionable behavior more harshly.

framing effects are rather commonplace, as options we choose from are always given within some kind of frame.[130]

6.2. What is there at stake with charity nudges?

As stipulated in the Introduction, the permissibility of charity nudges is discussed here for non-ideal circumstances. Whatever the character of duties towards the global poor may be, we can say with reasonable certainty that global institutions have so far employed sub-optimal methods in their attempts to alleviate global poverty. The task of improving these institutions is a persisting imperative in our normative considerations, but it is not at odds with an argument in favor of alleviating poverty through charity, or nudging for charity, in sub-optimal conditions. The discussion on the permissibility of charity nudges comes with a premonition that the shortcomings of global institutions in eliminating global poverty will not be overcome any time soon. An even stronger version of this position might be that we should be pessimistic about ever elevating the world's worst off via coercive measures. Even within single societies, Cohen claims, a perfect coercive structure that would result in maximal benefits for the worst off is unavailable, as "the vast economics literature on incentive-compatibility teaches that rules of the contemplated perfect kind cannot be designed" (1997, 10). The fear is then that if domestic institutions cannot guarantee the elimination of poverty, global institutions, which hold much less coercive power, can hope to do even less. I will work with the plausible assumption that the shortcomings in designing global institutions are fairly stable and persistent.

[130] For other notable studies in the area of charity nudges or with implications for how they may be designed, see Cialdini and Schroeder (1976) and Small et al. (2007).

Given that charity nudges seem to hold great potential for alleviating global poverty, and that resources for efficiently using charity nudges and recognizing people in most dire need will often only be available to big charity organizations, I also make the following assumptions: 1) charity programs using nudges will be more efficient than those which do not, and 2) big registered charities, like Oxfam and the Gates Foundation, will be more efficient in alleviating poverty than small charities and direct giving. These assumptions are highly contingent, and they may not constitute arguments against other forms of donation. However, they help us establish a baseline assumption about efficient charity giving for the purposes of this chapter.

The question then is whether citizens should endorse the use of charity nudges in their societies in the absence of effective domestic and global institutions. Even if we endorse charity nudges in non-ideal circumstances, we may still acknowledge that charity nudges correspondingly represent a less-than-ideal solution to global poverty, both in terms of effectiveness (compared to ideal global institutions) and moral desirability, but claim that they should be endorsed all things considered.

6.2.1. ***Charity nudges and personal autonomy***

The worry presented in this subsection is that while charity nudges may in fact be more effective than other means of tackling global poverty, they may interfere with the goals individuals want to pursue and how they want to organize their reflective capacities on a resource-based account of autonomy. The most comprehensive account of how charity nudges relate to autonomy is offered by Meena Krishnamurthy (2015). She suggests the following scenario: "A campaigner for a charity, C, comes to S's door and shows S a "persuasive pamphlet" that is designed to nudge S to donate to the charity that C is campaigning for. After giving S some time to peruse the pamphlet, C asks S to donate to the charity. S donates." (ibid., 253)

Is C violating S's autonomy by presenting the persuasive pamphlet?[131] Krishnamurthy believes we may interfere with one's autonomy if the influence we introduce moves one's actions away from the reasons "that stem from or relate to her own aims or commitments and not those of another individual" (ibid., 255). Enabling autonomy would entail we make sure that we do not divert individuals from acting consistently with aims and goals that they have set out for themselves. However, Krishnamurthy also seems to believe that for people to be autonomous, motivational drives have to emanate consistently from their aims and moral beliefs. In Krishnamurthy's words, "an individual is *autonomous* if she (1) has aims or commitments and (2) acts in ways that are motivated by these aims and commitments" (ibid., emphasis in original). People have an interest in their motivational drives corresponding with the strength of their reflected aims throughout their daily actions.

It is uncertain on Krishnamurthy's account how strong a bond is required between beliefs and motivations for action emanating from them, given our recent findings about behavioral influences and their pervasiveness. If the requirement is derived from strong internalism – that motivations and actions neatly follow aims and beliefs – then it is too demanding, as I have claimed in 5.1.2. Motivations are frequently driven by external factors, and Krishnamurthy in fact acknowledges that aims and motivations are often in a discrepancy. The differences in the accounts of various authors will often come down to the degree to which the drives of individuals can be external for those beliefs and commitments to be sincerely held. As I have claimed in Chapter 5, I will require, for the sincerity of moral belief or personal aim, that at least some motivation

[131] Krishnamurthy may have to acknowledge that pamphlets may be "persuasive" to varying degrees and that our normative judgments may vary in response to the strength of the pamphlet's "persuasiveness". Some nudges are easily resistible, and some are hardly resistible even if transparent. See my analysis in Chapter 3.

emanates from the belief, or that there is a fairly stable desire to act in line with the act or the belief, even if it does not always materialize in action.

My account of autonomy allows that people carry out many of their personal projects at low capacity, and invest limited reflective resources on goals that matter to them, in line with their own management styles. Non-reflective influences can thus be permissible in areas where individuals do not wish to invest reflective capacities, if such influences respect the internalist condition that they do not drive individuals away from their long-term plans and commitments. But even the investment of reflective resources can dynamically shift between the individual's projects. Specifically, even if individuals are motivated most of the time to act on one of their moral projects, it is completely regular that they have phases of "loosening up" to lead a psychologically stable autonomous life. This is either because some individuals value returning back on-course as part of building character, or they find it overly demanding and in fact stifling for their autonomy to be in a permanent mindset of a moral cause or personal project (see 5.4.3. again). It is with these considerations in mind that I turn to assessing Krishnamurthy's autonomy test for charity nudges.

Krishnamurthy suggests that S's moral beliefs and the motivations emanating from them might be exemplified in four character profiles. S might be a philanthropist, a weak-willed philanthropist, a misanthropist, and a neutralist. The philanthropist and the weak-willed philanthropist both believe they ought to help the poor, but the weak-willed philanthropist lacks the motivational drive to act upon his aim, and would not act unless he is nudged. The misanthropist does not aim to help the poor (nor is he motivated to do so). The neutralist is indifferent or undecided about helping the poor, and correspondingly lacks motivation that would arise from either aim (Krishnamurthy 2015, 256). We are to assume that S donates due to the

exposure to the pamphlet no matter which of these four profiles he happens to fit. What kind of person does S have to be, in order for his autonomy to be compromised by the persuasive pamphlet? We may try to get the seemingly straightforward cases out of the way first. The philanthropist, since he already aims to help the poor and acts from his own reasons, seems not to be violated by the pamphlet. The same reasoning would suggest that the autonomy of the misanthropist, whose donating is estranged from his own aims and commitments, is undermined.

But the philanthropist case might not be open-and-shut. Some might think that the addition of external influences changes the structure of the philanthropist's motivations even if his acts remain the same, and that, the claim goes, is sufficient for autonomy to be compromised. However, this suggestion fares badly with the pervasiveness stipulation, since we know individuals will be inevitably affected by many non-reflective influences. The objection also seems to subtly cling to strong internalism about moral motivations, expecting that the individual's motivational structure neatly flows from his moral beliefs. However, the very existence of weak-willed philanthropists brings this into question. If we allow for the existence of weak-willed philanthropists in the first place (and not claim that they are actually self-deluding or other-deluding misanthropists) then we immediately have to acknowledge that motivations often do not neatly flow from commitments. There are numerous causes why individuals fail to act on their commitments, and behavioral science finds many of these in systematic cognitive biases. We might consider it *regrettable* that our beliefs often cannot garner sufficient motivation, and that we need the help of nudges to act on our aims and commitments. But this is akin to regretting that akrasia or procrastination sometimes come in the way of acting on our moral duties. It would be too demanding for autonomy to depend

on the absence of such influences, given that they are a lasting evolutionary trait of the human condition.[132]

Let's turn now to the weak-willed philanthropist, whose pursuit of moral goals is in fact helped by charity nudges, under the counterfactual assumption that he would not act on his moral belief had he not been nudged. Krishnamurthy states that whether he is autonomous or not will depend on how much of his motivational drive for donating comes from the external source, in this case, the persuasive pamphlet (ibid., 257). There will be disagreement about how much is too much, but also, as Krishnamurthy says, how much drive the weak-willed philanthropist has in the first place. If he were to have no motivational drive emanating from his moral belief whatsoever or a fairly stable desire to help the poor, then, Krishnamurthy states, and I agree, he would not have a genuine aim to help the poor. To have a genuine aim involves at least some motivational drive on the weak-willed philanthropist's part (ibid.).[133] If weak-willed philanthropists do have some drive of their own, and do not mind pursuing some of their goals non-reflectively, the charity nudges facilitate the outcome-based condition of autonomy, helping them to boost motivation and act in accordance with their own aims and reasons. So long as S has some drive to help the poor, which confirms his background commitment, the reinforcement of the drive via charity nudges is compatible with S retaining autonomy.

It is possible, however, on a resource-based account of autonomy, that many (weak-willed) philanthropists not only want to invest reflective resources into acting on their moral goal, but that they want to act independently of some deliberate non-reflective influence of a nudger. These

[132] I thank Victor Tadros for helping me phrase this last point.
[133] Krishnamurthy also discusses the version of a weak-willed philanthropist who has no motivational drive to act on his aims, and claims that the nudge makes him non-autonomous (2015, 257-258), but she does seem to suggest, and I agree, that this psychological characterization of the weak-willed philanthropist is less feasible than the one I explicated.

philanthropists might believe that the bulk of their motivational drive has to come from their own internal struggle to carry out their aims, and that philanthropy is valuable only if their drive is not tampered by the external interference of nudgers. Of course, such individuals may overestimate their own cognitive capacities and may set out objectives for themselves that they cannot hope to carry out, but it is still reasonable to suggest that, from the standpoint of autonomy, charity nudges should not be used to interfere with their action.

With regard to the neutralist, who is undecided or indifferent about his aims of helping the poor, Krishnamurthy claims that influencing him makes his actions non-autonomous, since he is not acting from his own reasons (ibid., 258). But notice that on these conditions for autonomy, the neutralist would act non-autonomously no matter what he did, since his choice to act or not to act to help the poor would supposedly not correspond a specific aim. It seems more likely, on my view, that charity nudges do not interfere with the neutralist's autonomy, since he had no autonomous aims to interfere with.[134] They would be an interference only if they ran against a non-neutralist's aims, or if the neutralist adamantly commits to being undecided or indifferent. The latter case would be somewhat outlandish, regardless of whether the neutralist strives adamantly to maintain neutrality reflectively or at low capacity. Still, an account of autonomy that respects a diversity of management styles would require that even such idiosyncrasies be respected.

Let me shortly return to the misanthropist. There is some doubt about the feasibility of the misanthropist's character. If he acts on the influence of the pamphlet, he may not be fully committed to not helping the poor. Recall Saghai's point from 3.5.4. that nudgees are more likely to resist a nudge if it produces a feeling of dysfluency, and that this is more frequent when

[134] For a similar point in the area of cafeteria nudges, see Nys and Engelen (2017).

preferences are strong and settled (2013, 489). We should not expect the charity nudges described in 6.1.1. to completely dismantle the drives that normally lead us to action. If individuals are fully committed to certain aims, we would expect them to be more resilient to the effects of the non-reflective influences and less likely to cede to the effect. If not, then we should perhaps require that the described misanthropist in the scenario is at least a weak-willed misanthropist who sometimes cannot help himself but to donate to the poor. A weak-willed misanthropist would have his weakness abused and autonomy undermined by a charity nudge.

Considerations of autonomy for the case of charity nudges should also make us take note of some shortcomings in Krishnamurthy's scenario concerning her overly simplified character profiles. While people may indeed decide on a general commitment to help the poor through charitable donations, these aims are normally counterbalanced with other aims the person commits to. It is often the case, though sometimes not explicitly, that an individual endorses a rough ordering of these aims. The commitment to philanthropy, for instance, may be fully genuine and backed in motivational drive, but trumped by other aims. Furthermore, people have notions about when they have done enough for an aim at some given time. If S donates to charity in early October and feels he has done enough for that month, resisting a persuasive pamphlet in late October does not show that he is now a misanthropist. Rather, S decides on whether to act on his aims to help the poor in view of his other aims, which may hold temporary priority. What does this mean for charity nudges? It means, as the resource-based account indicates, that they could undermine the autonomies of agents in terms of the ways in which they want to pursue their aim sets. Charity nudges may disrupt the ordering of one's aims and commitments, or drive philanthropists into doing more than they want to do or feel that they owe. Hence, autonomy concerns about charity nudges do not only concern the weak-willed.

Let me recapitulate. Charity nudges may violate the autonomy of persons by driving them away from their aims or making them act in ways that are at odds with their strategies of pursuing their conceptions of the good.[135] Yet, charity nudges may also help people to autonomously act on unrealized philanthropic aims, as in the case of Krishnamurty's weak-willed philanthropists. In a consequentalist sense at least, charity nudges seem to be neutral towards autonomy, as they may both undermine some and enable the autonomies of others, while effectively aiding with the issue of global poverty. But this might not be enough for some critics. It is more important, the objection might go, not to violate autonomies than it is to enable them. The following sections offer multiple responses to this objection. The first is that the number of enabled autonomies may significantly outweigh the undermined, given that there are pervasive cognitive causes for people not acting on their aims to help the poor.

6.2.2. ***Biases against philanthropy***

In 5.1.1., primarily following Persson and Savulescu (2012), I detailed a number of psychological influences that are detrimental to the fulfilment of citizens' moral duties. Many of these are relevant for the duty of charity donation, in the face of ineffective institutions. For these reasons, many individuals will not adopt the moral duty of charity donation as part of their own aims and commitments. Yet, despite odds, others will in fact adopt a philanthropic attitude, but will then, under detrimental influence, fail to act on it in ways they would find appropriate. These negative effects will often be predictable to choice architects, and there will be little moral difference between them changing choice environments and leaving them as they were. Could

[135] That being said, the pursuit of one's autonomous goals, as I mention in the Introduction, could be morally worthless. I will later claim that misanthropists have no claim against charity nudges given at least the certainty that we have some obligations to help the global poor. Nevertheless, the self-government of misanthropists will be undermined, regardless of the fact that the undermining comes at no moral cost.

random, yet predictable choice environments be relevant for autonomy just as charity nudges? Are philanthropists' autonomies undermined by psychological biases against philanthropy?

Let's recall some of the detrimental influences from the previous chapter. An overly "localized" morality owed to our hunter-gatherer ancestry, the extension of care to those near and dear, an empathic insensitivity to rising numbers of sufferers, temporal shortsightedness – these all amount to what I will refer to here as a *proximity bias*. I understand it as the tendency of people to be influenced by the exposure (or a lack thereof) to certain social problems. Although individuals often have extensive information on the gravity of some moral issue, they are influenced in their actions by considerations of space, time, and their immediate social circles. An important related heuristic, which I have not yet considered, is availability, characterized as causing people "to assess the frequency of a class or the probability of an event by the ease with which instances or occurrences can be brought to mind" (Tversky and Kahneman 1974, 1127). As Tversky and Kahneman claim, there are different varieties to this heuristic, in which factors like the retrievability or the imaginability of instances play a role in decision-making processes. It is not that people are lacking information about what is beyond their attention span, but considerations about known information become much weightier when there is an experiential component, compared to being exposed to "troubling numbers" about the poor.

The fact that the plight of the poor is often experientially unavailable to those well-off may easily be contributing to the phenomenon of the "invisible poor" (Slovic 2010; Small and Loewenstein 2003). On the other hand, what is most accessible in people's decision-making are their personal projects – looking after their families and friends, doing their jobs, or sticking to routines. These considerations almost always enjoy the advantage of immediacy, and yet, many people, in retrospect, regret not donating more money to alleviate poverty.

I want to mention another psychological constraint at play – *just world beliefs* (JWB). Broadly, JWB entail that people's cognition "operate[s] under the assumption that the world is a just place, commonly expressed in the psychological literature as 'people get what they deserve and deserve what they get.' JWB are considered to be a real psychological phenomenon shared to some degree by all people" (Kasperbauer 2015, 218).[136] As Furnham shows in his research, the belief in a just world has adaptive functions for individuals (2003). It is evident that the coping mechanisms of sensing the world to be just and overlooking injustices stymie individual attempts at carrying out their aims to help the poor.

The effects of the proximity bias and just world beliefs can be expected to impede many individuals in their aim to help the poor. Charity nudges, on the other hand, can be expected to help these individuals in pressing for the realization of their aims appropriately within their sets of counterbalanced values. But this still may not be sufficient for critics. They could claim that even if the number of those whose autonomies are vitiated compared to those whose autonomies are enabled is small, it would still be wrong to pursue a more autonomous society on the backs of the vitiated. I now turn to the point that the autonomies of these individuals may not hold blanket priority given at least some certainty about moral duties towards the global poor.

6.2.3. ***Duty of poverty alleviation***

So far, this chapter has dealt with the permissibility of charity nudges by zooming in on their effects on personal autonomy. But we should also discuss the non-ideal circumstances of dealing with global poverty – there is broad consensus that global institutions are not doing nearly enough to allay the plight of the global poor. This means that our duties for alleviating poverty

[136] For more insights into 'just world beliefs', see Lerner (1980) and Rubin and Peplau (1975).

seem to re-emerge. Can charity nudges act as a supplementary tool for meeting our moral duties towards the poor?

An important problem about our duties to help the global poor is in an *uncertainty* about their exact character, or how they are grounded.[137] As I mentioned in 5.2., some duties are enforceable and feasibility-constrained non-enforceable. Such duties trump autonomy considerations. In other circumstances, duties are uncertain, either in the sense that we cannot decisively determine a duty all things considered, or in that we cannot determine how much wrong people have rights to do, or how much a duty requires given conflicting considerations.

The uncertainty is further reinforced by a lack of consensus among philosophers about what it is grounded in. Some, like Gillian Brock (2009), ground our duties of poverty alleviation in the underlying duty to treat all human beings with equal concern and respect, no matter where they happen to be. Others, like Peter Singer (2004), believe we have a duty to assist those in need if this is in our power, regardless of where they are, in the same way we have a duty to rescue a child drowning in a nearby pond. Yet, others, like Thomas Pogge (2005; 2007), argue that we have a duty to alleviate global poverty because of our causal responsibility for creating, upholding and benefiting from a global order that harms the poor.

The uncertainty about our duties for poverty alleviation is two-fold. On one hand, we have not been able to reach consensus on the exact grounding of the duty. On the other hand, and pertaining especially to the discussion herein, there is uncertainty regarding how the duty should be counterbalanced against a duty towards personal autonomy, and how the burdens are to be distributed. We may expect at least some variation with regard to the extent of our duty for poverty

[137] The uncertainty might contribute to the defectiveness of our global institutions, but I do not investigate this further here.

alleviation, when being counterbalanced against autonomy, depending on whether it is a duty of respect, a duty of humanity, or a duty of redress.[138] Consider the following case:

> "suppose Sue could either send her son to an excellent private school, or send him to an adequate state-funded school, enabling her to donate the money she would have spent to charities tackling global poverty. Let us imagine that Sue is uncertain whether it is permissible to send her son to the private school, given what else she could do with the money. However uncertain she may be about this matter, she either will or will not send her son to the private school." (Barry and Tomlin 2016, 898-899)

Krishnamurthy says that the permissibility of nudges will depend on solving the "broader debates in global ethics about the nature of our duties to aid the poor" (2015, 263). But this conclusion is overly cautious. It seems to be widely accepted, and far more *certain*, that we have at least *some* duties towards guaranteeing minimal subsistence to all human individuals. If we can determine that we certainly have *some* duties towards the global poor, and that domestic and global institutions fail to fulfill them, the duties once again go back to individuals. And if we assume that individuals do not have exemption rights against being included in the distribution of burdens, then they shirk their duties when they endorse a misanthropic attitude. I will now claim that misanthropists do not have a fair complaint against being interfered with by charity nudges.

6.3. The nudge ethos

The advocate of a nudge ethos for charity (NE) holds that duties of poverty alleviation justify the usage of a charity nudge package, the extensiveness of which is curbed by considerations of personal autonomy. The view starts with five assumptions: 1.) the efficacy of

[138] My intuition is that the duty of redress would likely imply the strongest moral duty, given the duty-bearer's supposed participation in causing the harm to the right-bearer.

charity nudging, 2.) the limitations of institutions, 3.) the undermining of autonomy of at least certain individuals by charity nudges, 4.) the existence of biases which deflect individuals from charitable donations, and 5.) a wide social commitment towards both poverty alleviation and personal autonomy.

The endorsers of the NE hold that institutional shortcomings should be compensated with charity nudges. This is not to say that charity nudging can or should do all the work in fulfilling a duty of poverty alleviation, but that it should complement the work of defective institutions. Individuals in the NE acknowledge their cognitive frailties, as well as those of their peers, in often being detracted from charity donations that they would otherwise want to make and to which they often commit. Their inabilities to act on their commitments may arise from status quo biases, proximity biases, just world beliefs, and many other lapses in reasoning.

My inspiration for the term 'nudge ethos' is drawn from Cohen's claim that justice is "not exclusively a function of its legislative structure, of its legally imperative rules, but also of the choices people make within those rules" (1997, 6). Cohen argues for an ethos that "informs individual choices" and "promotes a distribution more just than what the rules of the economic game by themselves can secure" (ibid., 10). Yet, Cohen acknowledges that we also require "agent-centered prerogatives" that allow individuals to follow, to an extent, their personal, rather than merely social goals (1992, 302-303). The nudge ethos similarly incorporates individual choices into the fulfillment of social justice, but while upholding the commitment to allow individuals to pursue their own personal projects. The NE is an ethos because the fulfillment of just conditions is procured through individual choices, and because individuals acknowledge charity nudges as driving their choices closer to optimally discharging their duties.

Given our duties towards the global poor, there is little reason not to interfere with a misanthropic attitude, and misanthropists have no right to complain about charity nudges. Endorsing the NE does, however, entail respect for how people aim to lead their lives and realize their personal projects. If the package of charity nudges is not overwhelming individuals, then they can be expected to respond to nudges differently given the current state of their personal affairs. Pressing personal matters and "bad timing" should build some momentary resilience to charity nudges, as individuals focus on issues that are more acute. These may involve crises in the pursuits of their goals, as well as helping friends, spouses, children, or family. At other times, these same individuals appreciate charity nudges, which make acting on their duties and commitments more likely. People often lose themselves in their projects even when they are committed to doing their part in helping the poor, and charity nudges aid in realizing these goals. It should not be assumed, however, that the urgency of personal matters builds a brick wall around agents and insulates them from the effects of charity nudges. Some individuals are more sensitive to the effects of nudges and may divert from their personal projects in critical situations. A commitment to autonomy thus requires that charity nudge packages are modest and that they can become transparent for individuals that are on the lookout for them, as I explained in Chapter 3. The NE thus allows the balance between the duty towards the poor and personal prerogatives to resolve itself spontaneously in the particular individual cases.

At the beginning of the chapter, I announced that nudging can help alleviate global poverty in the face of cooperation problems and information deficits. I explain each in turn. First, the NE informs individuals about a society-wide commitment that actions will be taken to meet collective duties. Michiru Nagatsu points to experiments on public goods claiming that “a substantial portion of subjects are *conditional* cooperators who cooperate if they expect that a sufficient number of

others do the same" (2015, 486). Thus, the commitments of misanthropists may well depend upon their expectations about the aims of others and whether these aims will be adequately pursued. As nudges have been proven powerful in boosting charity donations, conditional misanthropists may very well choose to cooperate. Following Bacharach (2006), Nagatsu also claims that the preferences of individuals may depend on whether they are faced with an I-frame or a we-frame (What should *I* do? vs. What should *we* do?) (2015, 487). When faced with an I-frame, a misanthropist chooses not to cooperate, but may opt for social cooperation and maximization of benefits when faced with a we-frame. The collective duty of poverty alleviation thus seems more likely to be consensual within the ethos, as an ethos faces individuals with a society-wide problem phrased in collective terms. Thus, as I mentioned in the previous chapter, a nudge project can help secure the conditions in which fulfilling collective duties becomes possible.

Second, I claim that the NE overcomes information deficits of a different kind. A common worry about charity programs is that even if everyone is a philanthropist, individual donations are scattered between various charity agencies. There is little combined effort to combat poverty in a strategic way. Individuals are often left to their own devices in figuring out which of the available options for donating is helpful to the global poor. Some failures of accumulating relevant data will be mistaken for a lack of willingness to do one's best in aiding the poor. The fear of information deficits should be somewhat downplayed with the NE, since individuals respond to charity nudges, that emphasize the options which refer back to findings by institutions or big charity organizations.

Granted, since I am stipulating that institutions are non-ideal, I should leave the possibility open that they could underperform in data accumulation and in drawing up effective strategies, and that less-than-adequate nudges could be designed as a result. I provide two responses for this worry. First, the endorsement of the NE does not entail that charity nudges must be implemented

across the board. Where there is skepticism about how cognitive biases affect us, how nudges work, or what benefits the poor, state institutions and charity agencies should show restraint in implementing nudges. The NE burdens institutions, charity organizations, nudge experts, and deliberating citizens with a lot of responsibility concerning the procurement of data before a nudge is implemented. Second, we should keep in mind that we are weighing between nudged and unregulated choice environments. Charity nudges might not be optimal for improving the position of the global poor, but they will still likely be superior to a choice environment where individuals lack coordination and are fettered by biases against philanthropy.

With this last point in mind, I want to reiterate that the NE is a non-ideal solution in non-ideal conditions. It is possible that the effects of charity nudges will, on occasion, produce discrepancies between people's motivational drives and their value sets (even when they are committed philanthropists). It is also possible that the nudge program that I am proposing will not capture the fulfillment of our duties and our commitments to goals and causes with perfect precision – it might go too far or not far enough. Non-ideal conditions will often require that we adopt solutions that are wrongful in some way, but are permissible all things considered.

Finally, I turn my attention to two possible versions of the NE. According to the *lax version*, nudges are meant only to facilitate philanthropic attitudes and alleviate global poverty. The charity nudges themselves are only meant to prompt changes in behavior, but not reflection about aims. A *strict version* of the NE, on the other hand, claims that a change in aims and attitudes through habituation is what should underlie a nudge program.[139] Although nudges predominantly affect people in non-reflective ways, the aftermath of nudged actions could prompt them into reflectively

[139] Once again, for remarks about habituation, see Bovens (2009).

adopting aims and principles that more correctly correspond to our social commitments about the global poor (or some other collective duty, as mentioned in Chapter 5). The strict view may even suggest that the lax view is incoherent in two ways. First, we cannot reflectively separate ourselves from behavioral outcomes. People are going to reflect on their behavior no matter what.[140] Second, behavioral experts like Carolyn will often be able to predict how behavior will impact reflection and changes in attitude, so a commitment against changing attitude seems untenable. The strict view thus endorses nudging as part of more permanent moral enhancement.

The lax view should not be principally opposed to changes in attitude if they were instigated merely by individuals reflecting more on the reasons for charitable giving that may become more salient within the NE. Yet, the lax view does not *aim directly* at changes in attitude, and opposes the institutions and charity organizations flooding citizens with nudges and thus interfering with their autonomies. The strict view, on the other hand, claims that the NE is already a non-ideal normative solution, and that the sacrifices the lax view makes should be reassessed. Why not sacrifice a bit more autonomy and overwhelm individuals with charity nudges if that can guarantee a more responsible citizenry that acts in unison? It remains open whether the distinction is tenable, if choice architects can reliably predict changes in attitude. For the moment, I will assume that they cannot. Reflection about behavior and the impact on moral behavior seems much less predictable than behavior at low capacity. As I showed in 2.2., it is with reference to System 2, most notably intelligence and reflective capacity (Evans 2009), that we should expect greater variation between persons. In the next subsection, I test both version of the nudge ethos against some notable objections.

[140] A reminder for why autonomy is not undermined specifically by reflective endorsement of nudged behavior can be found in 5.1.3.

6.3.1. ***Objections***

I tackle three important objections to the endorsement of the NE, in both the lax and the strict version.[141]

The first objection is that the NE contingently eventuates in notably inegalitarian outcomes. What makes them inegalitarian are the different burdens people take upon themselves in acting for the alleviation of global poverty. These inequalities emerge from two sources. First, the NE allows people to set their aims with regard to how much money, if any, they wish to spend on a charity program. In other words, it allows for the existence of misanthropists who, as I have established, act immorally by ignoring their share in the collective duty. Second, due to greater sensitivity to nudging effects, some people may find themselves making more donations than their personal projects can endure.[142] The force of this objection comes from the intuition that those who are engrossed in their own projects and are turning a deaf ear to duties towards the poor are in fact rewarded.

Although some inequality in taking up burdens is almost completely warranted, the strict view is not as affected by this objection, since its endorsers do not mind flooding the misanthropist with nudges in an attempt of changing his attitude. The lax view is vulnerable to the objection, but can offer the following responses. First, even if the lax NE accepts inegalitarian burdens, it does not have to accept that they are just; misanthropists do, in fact, act wrongly. Perhaps society should not show restraint in condemning free riders who allow their peers to take up all the burdens of personal contribution. Second, we should not lose sight of the inevitability we are faced with,

[141] Other objections to the nudge ethos could surely be raised about issues of moral compliance. However, seeing that I have given moral compliance very close attention in Chapter 5, and that charity nudges should not raise different compliance worries from other moral nudges, I leave these objections aside here.

[142] This worry, however, should be at least partly solved by ensuring nudge transparency of the kind I advocated in Chapter 3. We could imagine governments sharing the influences of charity organizations in a nudge registry.

which consists in a choice between the NE and an unregulated choice environment, where individuals are predictably biased towards not donating and where institutions fail to do their part. In both circumstances, we are faced with an inequality of burdens. Granted, the burdens might be greater for the philanthropists with the NE in place, but the NE also carries the benefits of discharging our duty towards the poor, while enabling the majority of people's autonomies. I claim that the inequality in burdens might be regrettable, but it is permissible all things considered. Third, I established in 5.2. that even in moral projects where the value that is being pursued clearly outweighs considerations of autonomy, there should be a constraint of proportionality. If the influences are such that they deeply disturb mental lives, or individuals are affected to the extent that they are brought on the brink of material destitution, then we have reasons to curb the moral nudge project by avoiding certain nudges or nudge stacking.

This response, however, suggests that society will have to accept inegalitarian burdens as a side-effect of the NE. This brings me to the second objection. If society will have to put up with misanthropists, who fail to do their part of the share, David Miller asks whether others are required to take up the slack (2011). In the context of charity nudges, the question is whether abiding individuals (philanthropists) should be exposed to more nudging to compensate for the slacking of misanthropists. Miller says the question does not apply to an "altruistic prisoner's dilemma" where the responsibilities of individuals are indeterminate (ibid., 231), but where it is rather "clear what justice requires each person to do by way of contribution" (ibid., 232). I did suggest earlier that the duties of poverty alleviation are indeterminate, but I claimed that there is also certainty about their existence, and the fact that we are not doing enough to discharge them, given ineffective institutions. I could claim that there are determinate requirements of poverty alleviation at least up to a certain point and that 'the slack' refers to these requirements. So, should philanthropists take

up the slack? Following Miller, three normative solutions are available: 1.) philanthropists can be expected to do as much as they were expected originally despite the misanthropist's non-abidance; 2.) they are required to do more; 3.) they are allowed to do less (ibid., 233-234). If solution 2 is correct, then exposing philanthropists to more charity nudging than the modest nudge package might be permissible.

Solution 3 should be unappealing for both the lax and the strict view. As Miller rightly points out, it only holds appeal if non-abidance is so prevalent that little can be achieved by only philanthropists helping the poor, or if non-abidance is not an injustice to the victims, but places those who abide at a disadvantage (ibid., 238). With regard to the first circumstance, donating to charity standardly produces a lot of good even if some people in society slack. There is, of course, tragedy in not being able to elevate all of those affected by poverty over a threshold of survival, but that is hardly an argument against elevating only some people when this is what resources allow. With regard to the second circumstance, non-abidance certainly is an injustice – it is what drives the argument. It may well happen that some individuals find themselves at a disadvantage compared to the misanthropists. This is partly why the lax view argues for a contained nudge project which allows individuals to balance their duties with their personal projects. But allowing individuals to withhold their help entirely because of other people's neglect amounts to tolerating a denial of a duty, which is unacceptable.

Choosing between solutions 1 and 2 is more difficult. While duties of alleviating poverty are indeterminate, it is certain we have some duties in a world with defective global institutions. In other words, while we do not know *how much* we are supposed to do, we know we ought to do *at least something*, and that 'something' amounts to a feasibility-constrained non-enforceable duty. Yet, there is little consensus on how much 'something' is. Let us imagine, however, that there is

such a line, even if we cannot tell exactly where to draw it. This would entail that when people do nothing, they obviously neglect a feasibility-constrained non-enforceable duty. Should we pick up other people's duties if they completely fail to act on them? While appropriating shares in duties might not seem all that bad when only some shirk their duty, it seems to lead to a disturbing implication. Imagine that a vast majority in a society fails to act on the duty, and only a handful of philanthropists remain willing. It seems wrong to say that the aggregate of neglected duties is now transferred to the philanthropist group and that they are at fault when they fail to discharge it. Feasibility-constrained non-enforceable duties do not seem to be transferable in this way, even if people (like the philanthropist group) find good reasons to discharge them. Endorsers of both the lax and the strict view should commit to the strategy according to which no slack should be taken up (solution 1). Both views ought to hold that we should alleviate global poverty as much as possible, but not that duties should be transferable in the way I just described. It could be claimed that the strict view proponent may want to adopt taking up the slack, since having more duties might encourage people to push their commitment to philanthropy higher up their priority list. Yet, this might be going too far even for the strict view proponent. The desired effect, he might say, is that the NE changes misanthropic attitudes, not that it turns duty-abiding people into saints.

I turn to one final objection. Does implementing a nudge ethos jeopardize people's capacities to reflect on their duties and commitments towards the global poor? Evan Riley argues nudges may cause reflective incapacitational injustice, which "is reflected in failures to support the development and(or) exercise of our reflective capacities for critical reasoning" in worlds like ours (2017, 599). The worry is that because nudge programs fail to support reflective development, they might be epistemically unjust. The phrase 'worlds like ours' is not merely a figure of speech – it refers to Sunstein's claim that a program of beneficent nudges is morally justified in a world

like ours (a world with behavioral failures) (Sunstein 2015b). It is difficult to establish whether a society with a nudge ethos, one that I have been describing, is too far removed from 'a world like ours' to be in the crosshairs of Riley's criticism. It would appear at least, given that individuals in the NE are 'people like us', that it is not.

I provide two reasons for why the worry is overstated. First, the burden of proof that behavioral effects do (not) jeopardize the development of critical reasoning should be both on the advocates as well as the critics of behavioral techniques. Commonsensically, the fact that we have been influenced by a charity nudge does not mean that the act of donating blows us by. If so, people who act on charity nudges can reflect on their actions and may conceivably evoke reasons to donate more vividly than people who have not been exposed or did not act on charity nudges. Thus, the judgment that charity nudges make people act on their duties more robotically than they otherwise would is presumptuous. Second, my assumptions about the nudge ethos came with certain safeguards. With the ethos in place, I assumed individuals to be aware of each other's cognitive deficiencies and the resources that may help them avoid nudges with which they disagree. I have also assumed a general society-wide debate on requirements of poverty and a charity nudge package that may help meet them. I do not believe that these assumptions are too outlandish, and thus sever ties with worlds like ours. These two reasons are enough to take most force from Riley's criticism.

Let me take stock. In this chapter, I have been scrutinizing the character of charity nudges and assessing their permissibility against a background of conflicting values. I found that charity nudges can indeed vitiate the autonomy of certain individuals, although there are good reasons to suggest that they enable the autonomies of most others. Charity nudges also show promise in helping us to act on our feasibility-constrained non-enforceable duties towards the global poor.

For these reasons, I have argued that promoting a *nudge ethos* is permissible all things considered. A nudge ethos strikes a balance between our duties towards the global poor and respecting people's autonomies. It also helps people to act upon their philanthropic attitudes, coordinate their efforts, and overcome information deficits. Although some objections can be raised, I believe the reasons for adopting charity nudges are overwhelming.

Conclusion

My aim in this dissertation was to explore whether the practice of nudging can be reconciled with the basic political principles of a liberal democracy, as well as to show that a serious consideration of recent strides in cognitive and behavioral science about reflexive and automatic cognition open up a wider range of normative questions and avenues than pursued thus far. This broader theoretical approach – which I labelled behavioral enhancement – looks not only at whether governments nudging citizens is permissible for advancing welfarist purposes, but also whether various kinds of non-reflective influences, from different sources, are permissible against the backdrop of liberal principles. In the dissertation specifically, I focus on investigating how we should cultivate and regulate pervasive behavioral influence, paying special attention to whether autonomy is protected, and harm is prevented or alleviated.

I believe that I have successfully laid the foundations of behavioral enhancement in this dissertation, although in more than one sense, the exploration remains incomplete. Let me say a few words about the dissertation's successes. First, the broader approach helps in solving the conundrums of the standard nudge debate. I have claimed that with a specific set of institutional provisions in place – which I called 'watchfulness' – nudges are a permissible and desirable boost to individual pursuits, allowing those who agree with the nudge to save up their reflective resources, and others to go their separate ways. Watchfulness, I argued, consists of several conditions that enable the dissenters to anticipate and see the influences that they disagree with, in order to circumvent them. Importantly, my institutional solution disposes of the challenge I detail in the Introduction that nudges cannot be democratic and useful, all at once.

Second, I show that worries about nudges undermining autonomy should in fact raise alarms about whether autonomy is under more serious threat from the non-reflective influences of

marketers. This is because marketers' influences are more likely to overwhelm agents, and are not curtailed by principles of government nudging, which pertain to mildness and respecting the nudgee's preferences. Although the market is normally considered to be part of the basic structure on Rawlsian accounts, the consideration of market influences has thus far been widely neglected both in the ethics of nudging and moral and political philosophy more broadly. I show that if behavioral influences have the capacity to undermine autonomy, the market setting will either require strong justification to keep its practices or will have to undergo considerable regulation. This broadens the nudge debate by rejecting the standard normative relationship – between government as the influencer, and citizen as the influencee.

Third, I show that non-reflective influences can be used for purposes other than the advancement of individual well-being. We can instead turn our attention to utilizing nudges to induce behavior that supports other values. In some cases, this seems hardly controversial, given the obvious priority of, say, alleviating grave harm in traffic compared to the value of personal autonomy; there will be little, if any value in ensuring conditions of transparency and democratic contestation. In other cases, it will not be obvious what duties individuals have, although it will be certain that they should do more than nothing; these will require curtailments of a carefully developed ethos. Some issues will also be raised with *how* these other-regarding considerations are pursued – whether they undermine the engagement with moral reasons that we often take to be at the core of truly moral lives, and whether they eradicate moral disagreement vital for moral progress. I claimed that both fears are exaggerated, and that moral nudges normally make moral compliance and moral disagreement no worse than they would have been in the absence of nudges. Future research will show how nudges morally compare to other methods of moral enhancement.

Although my efforts here help in determining the permissibility of influences more widely conceived, there are very important areas of behavioral enhancement still left unexplored. I have hinted at some throughout the dissertation. For instance, a democratic theorist might wonder whether politicians should, in any way, be restricted with regard to how they use non-reflective influences when they compete for public office, or how the political landscape should be otherwise behaviorally regulated. We might believe governments should aspire for behavioral neutrality (when possible) when setting up a political agenda or posing a referendum question. A social theorist might be interested in whether activists should be permitted to drive their causes on the backs of nudges. Furthermore, the case of charity nudges gave us an insight into what will often be at stake in debates about the permissibility of influences promoting other-regarding purposes, but these might not perfectly translate into debates about using nudges to oust discrimination, to promote social justice, or to ensure equality of opportunity. Still, I establish the philosophical foundation in this dissertation for unexplored debates on behavioral enhancement by offering a psychologically-informed account of personal autonomy; this will be the starting point for most such future debates.

I also invested great effort in the dissertation to show that my account of permissible behavioral influence is compatible with liberal principles, especially in view of the theoretical foundations I have listed in the Introduction. Whether the account and the whole project of behavioral enhancement can keep its liberal credentials will undoubtedly be contested. For example, it might be claimed that I fail to empirically update other liberal values – like non-discrimination, responsibility, or equality – in the way that I have updated the principle of personal autonomy, and that I thus fail to give these other central liberal values proper consideration. I believe this criticism is true, although I could have hardly done more within the span of the

dissertation. Hence, a more thorough assessment of whether political liberalism is compatible with behavioral enhancement will have to wait on a more complete elaboration of both.

I finish on an optimistic note. I believe that ethics of nudging has turned our attention to relevant scientific insights, which will, in time, help us to rethink and reconfigure some of the central notions of moral philosophy. Albeit far from being the final word on the subject, this dissertation serves the purposes of easing the transition and lighting the beacon to other similar philosophical explorations, still to come.

Bibliography

Abdukadirov, Sherzod. 2016. Who Should Nudge? In *Nudge Theory in Action: Behavioral Design in Policy and Markets*, ed. Sherzod Abdukadirov, 159–191. Palgrave Macmillan.

Akerlof, George A. 1991. Procrastination and Obedience. *The American Economic Review* 81 (2): 1–19.

——, and Robert J. Shiller. 2015. *Phishing for Phools: The Economics of Manipulation and Deception*. Princeton: Princeton University Press.

Allcott, Hunt. 2011. Social Norms and Energy Conservation. *Journal of Public Economics* 95 (9-10): 1082–1095.

Amir, On, and Dan Ariely. 2007. Decisions by Rules: The Case of Unwillingness to Pay for Beneficial Delays. *Journal of Marketing Research* 44 (1): 142–152.

Anderson, Joel. 2010. Review of Nudge: Improving Decisions about Health, Wealth, and Happiness. *Economics and Philosophy* 26 (3): 369–376.

Arad, Ayala, and Ariel Rubinstein. 2018. The People's Perspective on Libertarian-Paternalistic Policies. *The Journal of Law and Economics* 61 (2): 311–333.

Ariely, Dan. 2008. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. New York: HarperCollins Publishers.

——, and Klaus Wertenbroch. 2002. Procrastination, Deadlines, and Performance: Self-Control by Precommitment. *Psychological Science* 13 (3): 219–224.

Arneson, Richard J. 2000. Perfectionism and Politics. *Ethics* 111 (1): 37–63.

Ashcroft, Richard E. 2011. Personal financial incentives in health promotion: Where do they fit in an ethic of autonomy? *Health Expectations* 14 (2): 191–200.

Bacharach, Michael. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*, eds. Natalie Gold, and Robert Sugden. Princeton: Princeton University Press.

Baldwin, Robert. 2014. From Regulation to Behaviour Change: Giving Nudge the Third Degree. *The Modern Law Review* 77 (6): 831–857.

Bar-Gill, Oren. 2004. Seduction by Plastic. *Northwestern University Law Review* 98 (4): 1373–1434.

Barry, Christian, and Patrick Tomlin. 2016. Moral uncertainty and permissibility: Evaluating Option Sets. *Canadian Journal of Philosophy* 46 (6): 898–923.

Barton, Adrien, and Till Grüne-Yanoff. 2015. From Libertarian Paternalism to Nudging–and Beyond. *Review of Philosophy and Psychology* 6 (3): 341–359.

Bearak, Max. 2017. Hungary accused of 'hatemongering' in national survey targeting George Soros. *www.washingtonpost.com* (November 8, 2017). Available at: https://www.washingtonpost.com/news/worldviews/wp/2017/11/08/hungary-accused-of-hatemongering-in-national-survey-targeting-george-soros/.

Beggs, Jodi N. 2016. Private-Sector Nudging: The Good, the Bad, and the Uncertain. In *Nudge Theory in Action: Behavioral Design in Policy and Markets*, ed. Sherzod Abdukadirov, 125–158. Palgrave Macmillan.

Beraldo, Sergio. 2017. An Impossibility Result on Nudging Grounded in the Theory of Intentional Action. *CSEF Working Papers* 485.

Berlin, Isaiah. 1969. *Four Essays on Liberty*. Oxford: Oxford University Press.

Birks, David, and Alena Buyx. 2018. Punishing Intentions and Neurointerventions. *American Journal of Bioethics Neuroscience* 9 (3): 133–143.

Blumenthal-Barby, Jennifer S. 2012. Between Reason and Coercion: Ethically Permissible Influence in Health Care and Health Policy Contexts. *Kennedy Institute of Ethics Journal* 22 (4): 345–366.

——. 2013. Choice Architecture: A mechanism for improving decisions while preserving liberty? In *Paternalism: Theory and Practice*, eds. Christian Coons, and Michael Weber, 178–196. New York: Cambridge University Press.

——. 2016. Biases and Heuristics in Decision Making and Their Impact on Autonomy. *The American Journal of Bioethics* 16 (5): 5–15.

——, and Aanand D. Naik. 2015. In Defense of Nudge–Autonomy Compatibility. *The American Journal of Bioethics* 15 (10): 45–47.

Bohnet, Iris, Alexandra van Geen, and Max Bazerman. 2015. When Performance Trumps Gender Bias: Joint vs. Separate Evaluation. *Management Science* 62 (5): 1225–1234.

Bostrom, Nick, and Toby Ord. 2006. The Reversal Test: Eliminating Status Quo Bias in Applied Ethics. *Ethics* 116 (4): 656–679.

Bovens, Luc. 2009. The Ethics of Nudge. In *Preference Change: Approaches from Philosophy, Economics and Psychology*, eds. Till Grüne-Yanoff, and Sven Ove Hansson, 207–219. Berlin & New York: Springer.

Bradshaw, Della. 2015. How a little nudge can lead to better decisions. *www.ft.com* (November 15, 2015). Available at: https://www.ft.com/content/e98e2018-70ca-11e5-ad6d-f4ed76f0900a.

Brock, Gillian. 2009. *Global Justice: A Cosmopolitan Account.* Oxford, New York: Oxford University Press.

Brownstein, Michael. 2015. Implicit Bias. *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Available at: https://plato.stanford.edu/entries/implicit-bias/.

Bublitz, Jan Christoph, and Reinhard Merkel. 2014. Crimes Against Minds: On Mental Manipulations, Harms and a Human Right to Mental Self-Determination. *Criminal Law and Philosophy* 8 (1): 51–77.

Buchanan, Allen E. 1985. *Ethics, Efficiency, and the Market*. Oxford: Oxford University Press.

Buehler, Roger, Dale W. Griffin, and Michael Ross. 1994. Exploring the "Planning Fallacy": Why People Underestimate Their Task Completion Times. *Journal of Personality and Social Psychology* 67 (3): 366–381.

Burgess, Adam. 2012. 'Nudging' Healthy Lifestyles: The UK Experiments with the Behavioural Alternative to Regulation and the Market. *European Journal of Risk Regulation* 3 (1): 3–16.

Buss, Sarah. 2012. Autonomous Action: Self-Determination in the Passive Mode. *Ethics* 122 (4): 647–691.

Bykvist, Krister. 2017. Moral Uncertainty. *Philosophy Compass* 12 (3): 1–8.

Cabinet Office Behavioural Insights Team and The Charities Aid Foundation. 2012. Applying behavioural insights to reduce fraud, error and debt. Available at: http://38r8om2xjhhl25mw24492dir.wpengine.netdna-cdn.com/wp-content/uploads/2015/07/BIT_FraudErrorDebt_accessible.pdf.

——. 2013. Applying behavioural insights to charitable giving. Available at: http://good2give.ngo/wp-content/uploads/2016/04/CAF_Charitable_Giving_Report_May_2013.pdf.

Chandon, Pierre, J. Wesley Hutchinson, Eric T. Bradlow, and Scott H. Young. 2009. Does In-Store Marketing Work? Effects of the Number and Position of Shelf Facings on Brand Attention and Evaluation at the Point of Purchase. *Journal of Marketing* 73 (6): 1–17.

Chou, Eileen Y., and J. Keith Murnighan. 2013. Life or Death Decisions: Framing the Call for Help. *PLoS ONE* 8 (3).

Christman, John. 1988. Constructing the Inner Citadel: Recent Work on the Concept of Autonomy. *Ethics* 99 (1): 109–124.

——. 1991a. Liberalism and Individual Positive Freedom. *Ethics* 101 (2): 343–359.

——. 1991b. Autonomy and Personal History. *Canadian Journal of Philosophy* 21 (1): 1–24.

Cialdini. Robert B., and David A. Schroeder. 1976. Increasing Compliance by Legitimizing Paltry Contributions: When Even a Penny Helps. *Journal of Personality and Social Psychology* 34 (4): 599–604.

Clayton, Matthew, and David Stevens. 2015. Is the Free Market Acceptable to Everyone? *Res Publica* 21 (4): 363–382.

——, and Andrés Moles. 2018. Neurointerventions, Morality, and Children. In *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*, eds. David Birks, and Thomas Douglas, 235–251. Oxford: Oxford University Press.

Cohen, Gerald A. 1992. Incentives, Inequality, and Community. In *The Tanner Lectures on Human Values, Volume Thirteen*, ed. Grethe B. Peterson, 261–329. Salt Lake City: University of Utah Press.

——. 1997. Where the Action is: On the Site of Distributive Justice. *Philosophy and Public Affairs* 26 (1): 3–30.

——. 2013. *Finding Oneself in the Other* (edited by Michael Otsuka). Princeton and Oxford: Princeton University Press.

Cohen, Shlomo. 2015. A Philosophical Misunderstanding at the Basis of Opposition to Nudging. *The American Journal of Bioethics* 15 (10): 39–41.

Conly, Sarah. 2012. *Against Autonomy: Justifying Coercive Paternalism*. New York: Cambridge University Press.

Coons, Christian, and Michael Weber. 2013. Introduction: Paternalism – Issues and Trends. In *Paternalism: Theory and Practice*, eds. Christian Coons, and Michael Weber, 1–24. New York: Cambridge University Press.

Dasgupta, Nilanjana. 2013. Implicit Attitudes and Beliefs Adapt to Situations: A Decade of Research on the Malleability of Implicit Prejudice, Stereotypes, and the Self-Concept. *Advances in Experimental Social Psychology* 47: 233–279.

DeGrazia, David. 2012. Moral enhancement, freedom, and what we (should) value in moral behaviour. *Journal of Medical Ethics* 40 (6): 361–368.

DellaVigna, Stefano, and Ulrike Malmendier. 2006. Paying Not to Go to the Gym. *American Economic Review* 96 (3): 694–719.

Dennett, Daniel C. 1969. *Content and Consciousness*. London: Routledge and Kegan Paul.

Desvouges, William H., F. Reed Johnson, Richard W. Dunford, Kevin J. Boyle, Sara P. Hudson, and K. Nicole Wilson. 1993. Measuring Natural Resource Damages with Contingent Valuation: Tests of Validity and Reliability. In *Contingent Valuation: A Critical Assessment*, ed. Jerry A. Hausman, 91–164. New York: North-Holland.

Dinner, Isaac, Eric J. Johnson, Daniel G. Goldstein, and Kaiya Liu. 2011. Partitioning Default Effects: Why People Choose Not to Choose. *Journal of Experimental Psychology: Applied* 17 (4): 332–341.

Double, Richard. 1992. Two Types of Autonomy Accounts. *Canadian Journal of Philosophy* 22 (1): 65–80.

Douglas, Thomas. 2013. Moral enhancement via direct emotion modulation: A reply to John Harris. *Bioethics* 27 (3): 160–168.

——. 2018. Neural and Environmental Modulation of Motivation. In *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*, eds. David Birks, and Thomas Douglas, 208–224. Oxford: Oxford University Press.

——. Unpublished manuscript. The 'Mere Substitution' Defense of Nudging, and its Extension to Neurointerventions.

Dworkin, Gerald. 1976. Autonomy and Behavior Control. *The Hastings Center Report* 6 (1): 23–28.

——. 1981. The Concept of Autonomy. *Grazer Philosophische Studien* 12 (1): 203–213.

——. 1988. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.

——. 2002. Paternalism. *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Available at: https://plato.stanford.edu/entries/paternalism/.

Egebark, Johan, and Mathias Ekström. 2013. Can Indifference Make the World Greener? *IFN Working Paper* 975.

Elster, Jon. 1995. Strategic Uses of Argument. In *Barriers to Conflict Resolution*, eds. Kenneth Arrow, Robert H. Mnookin, Lee Ross, Amos Tversky, and Robert Wilson, 237–257. New York: Norton.

Engelen, Bart, Alan Thomas, Alfred Archer, and Niels van de Ven. 2018. Exemplars and nudges: Combining two strategies for moral education. *Journal of Moral Education* 47 (3): 346–365.

Epstein, Richard A. 2006. Behavioral Economics: Human Errors and Market Corrections. *University of Chicago Law Review* 73 (1): 111–132.

Estlund, David. 2014. Utopophobia. *Philosophy and Public Affairs* 42 (2): 113–134.

European Parliament, and European Council. 2003. Directive 2003/33/EC on the approximation of the laws, regulations and administrative provisions of the Member States relating to the advertising and sponsorship of tobacco products (Text with EEA relevance). *Official Journal of the European Union* L 152.

——. 2018. Regulation (EU) 2018/1725 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC. *Official Journal of the European Union* L 295/39.

Evans, Jonathan St.B.T. 2009. How many dual-processes do we need? One, two, or many? In *In Two Minds: Dual Processes and Beyond*, eds. Jonathan St.B.T. Evans, and Keith Frankish, 33–54. Oxford: Oxford University Press.

——, and David E. Over. 1996. *Rationality and Reasoning*. Psychology Press, Hove.

Farrell, Henry. 2017. This year's economics Nobel winner invented a tool that's both brilliant and undemocratic. *www.vox.com* (October 16, 2017). Available at: https://www.vox.com/the-big-idea/2017/10/16/16481836/nudges-thaler-nobel-economics-prize-undemocratic-tool.

Feinberg, Joel. 1986a. *Harm to Self*. New York, Oxford: Oxford University Press.

——. 1986b. *Harm to Others*. New York, Oxford: Oxford University Press.

Felsen, Gidon, Noah Castelo, and Peter B. Reiner. 2013. Decision enhancement and autonomy: Public attitudes towards overt and covert nudges. *Judgment and Decision Making* 8 (3): 202–213.

Foot, Philippa. 1967. The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review* 5: 5–15.

Frankfurt, Harry. 1969. Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy* 66 (23): 829–839.

——. 1988. *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Frankish, Keith, and Jonathan St.B.T. Evans. 2009. The duality of mind: An historical perspective. In *In Two Minds: Dual Processes and Beyond*, eds. Jonathan St.B.T. Evans, and Keith Frankish, 1–32. Oxford: Oxford University Press.

——. 2009. Systems and levels: Dual-system theories and the personal–subpersonal distinction. In *In Two Minds: Dual Processes and Beyond*, eds. Jonathan St.B.T. Evans, and Keith Frankish, 89–108. Oxford: Oxford University Press.

Friedman, Milton. 1982. *Capitalism and Freedom*, 2nd edition. Chicago: The University of Chicago Press.

Furnham, Adrian. 2003. Belief in a Just World: Research Progress over the Past Decade. *Personality and Individual Differences* 34 (5): 795–817.

Gazzaniga, Michael S. 1998. *The Mind's Past*. Berkeley: University of California Press.

George, David. 2001. *Preference Pollution: How Markets Create the Desires We Dislike*. Ann Arbor: The University of Michigan Press.

Gigerenzer, Gerd. 2007. *Gut Feelings*. London: Penguin.

——. 2014. *Risk Savvy: How to Make Good Decisions*. London: Penguin.

Gilovich, Thomas, Dale Griffin, and Daniel Kahneman (eds.). 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.

Gino, Francesca, and Sreedhari D. Desai. 2012. Memory Lane and Morality: How Childhood Memories Promote Prosocial Behavior. *Journal of Personality and Social Psychology* 102 (4): 743–758.

Grill, Kalle. 2007. The Normative Core of Paternalism. *Res Publica* 13 (4): 441–458.

——. 2014. Expanding the Nudge: Designing Choice Contexts and Choice Contents. *Rationality, Markets and Morals* 5: 139–162.

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

——. 2016. Why Behavioural Policy Needs Mechanistic Evidence. *Economics and Philosophy* 32 (3): 463-483.

——, and Ralph Hertwig. 2016. Nudge Versus Boost: How Coherent Are Policy and Theory? *Minds and Machines* 26 (1-2): 149-183.

Hagman, William, David Andersson, Daniel Västfjäll, and Gustav Tinghög. 2015. Public Views on Policies Involving Nudges. *Review of Philosophy and Psychology* 6 (3): 439–453.

Hansen, Pelle G., and Andreas M. Jespersen. 2013. Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy. *European Journal of Risk Regulation* 4 (1): 3–28.

Harris, John. 2011. Moral Enhancement and Freedom. *Bioethics* 25 (2): 102–111.

Hauskeller, Michael. 2013. *Better Humans?* Durham: Acumen Publishing.

Hausman, Daniel M., and Brynn Welch. 2010. Debate: To Nudge or Not to Nudge. *Journal of Political Philosophy* 18 (1): 123–136.

Heilmann, Conrad. 2014. Success conditions for nudges: a methodological critique of libertarian paternalism. *European Journal of Philosophy of Science* 4 (1): 75–94.

Henderson, John M., and Andrew Hollingworth. 1999. High-level Scene Perception. *Annual Review of Psychology* 50 (1): 243–271.

Herstein, Ori J. 2012. Defending the Right to Do Wrong. *Law and Philosophy* 31 (3): 343–365.

House of Lords, Science and Technology Select Committee. 2011. Behaviour Change: Report (HL Paper 179). Available at: https://publications.parliament.uk/pa/ld201012/ldselect/ldsctech/179/179.pdf.

Hull, David L. 2001. *Science and Selection: Essays on Biological Evolution and the Philosophy of Science*. Cambridge: Cambridge University Press.

Institute for Government and the Cabinet Office. 2010. MINDSPACE: Influencing behaviour through public policy. Available at: https://www.instituteforgovernment.org.uk/sites/default/files/publications/MINDSPACE.pdf .

Ivanković, Viktor. 2016. Christiano's Deliberative Expertism and Choice Architecture. *Annals of the Croatian Political Science Association* 12 (1): 83–101.

——, and Bart Engelen. 2019. Nudging, Transparency, and Watchfulness. *Social Theory and Practice* 45 (1): 43–73.

Jefferson, Anneli, Lisa Bortolotti, and Bojana Kuzmanovic. 2017. What is unrealistic optimism? *Consciousness and Cognition* 50: 3–11.

John, Peter. 2018. *How Far to Nudge? Assessing Behavioural Public Policy*. London: Elgar.

——, Graham Smith, and Gerry Stoker. 2009. Nudge Nudge, Think Think: Two Strategies for Changing Civic Behaviour. *The Political Quarterly* 80 (3): 361–370.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. London: Penguin.

——, Paul Slovic, and Amos Tversky (eds.). 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

——, Jack L. Knetsch, and Richard H. Thaler. 1991. Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *The Journal of Economic Perspectives* 5 (1): 193–206.

——, and Amos Tversky. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5 (4): 297–323.

——, Ilana Ritov, Karen E. Jacowitz, and Paul Grant. 1993. Stated Willingness to Pay for Public Goods: A Psychological Perspective. *Psychological Science* 4 (5): 310–315.

——, and Amos Tversky (eds.). 2000. *Choices, Values, and Frames*. Cambridge: Cambridge University Press.

Kamin, Kim A., and Jeffrey J. Rachlinski. 1995. Ex Post ≠ Ex Ante: Determining Liability in Hindsight. *Law and Human Behavior* 19 (1): 89–104.

Karlan, Dean, and John A. List. 2007. Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment. *American Economic Review* 97 (5): 1774–1793.

Kasperbauer, Tyler J. 2015. Psychological Constraints on Egalitarianism: The Challenge of Just World Beliefs. *Res Publica* 21 (3): 217–234.

——. 2017. The permissibility of nudging for sustainable energy consumption. *Energy Policy* 111 (C): 52–57.

Kelly, Jamie T. 2012. *Framing Democracy: A Behavioral Approach to Democratic Theory*. Princeton: Princeton University Press.

——. 2013. Libertarian paternalism, utilitarianism, and justice. In *Paternalism: Theory and Practice*, eds. Christian Coons, and Michael Weber, 216–230. New York: Cambridge University Press.

Keysar, Boaz, Sayuri L. Hayakawa, and Sun Gyu An. 2012. The Foreign-Language Effect: Thinking in a Foreign Tongue Reduces Decision Biases. *Psychological Science* 23 (6): 661–668.

Korsgaard, Christine M. 2009. *Self-Constitution: Agency, Identity, and Integrity.* Oxford: Oxford University Press.

Krishnamurthy, Meena. 2015. Nudging Global Poverty Alleviation? *Law and Ethics of Human Rights* 9 (2): 249–264.

Krueger, Joachim I., and David C. Funder. 2004. Towards a balanced social psychology: Causes, consequences and cures for the problem-seeking approach to social cognition and behavior. *Behavioral and Brain Sciences* 27 (3): 313–327.

Laborde, Cecile. 2017. *Liberalism's Religion*. Oxford: Oxford University Press.

Lades, Leonhard K. 2014. Impulsive consumption and reflexive thought: Nudging ethical consumer behavior. *Journal of Economic Psychology* 41: 114–128.

LeBoeuf, Robyn A., and Eldar Shafir. 2003. Deep Thoughts and Shallow Frames: On the Susceptibility to Framing Effects. *Journal of Behavioral Decision Making* 16 (2): 77–92.

Lepenies, Robert, and Magdalena Małecka. 2015. The Institutional Consequences of Nudging – Nudges, Politics, and the Law. *Review of Philosophy and Psychology* 6 (3): 427–437.

Lerner, Melvin J. 1980. *The Belief in a Just World: A Fundamental Delusion*. New York: Springer.

Lindstrom, Martin. 2008. *Buyology: Truth and Lies About What We Buy*. New York: Doubleday.

——. 2011. *Brandwashed: Tricks Companies Use to Manipulate Our Minds and Persuade Us to Buy*. New York: Crown Business.

Lockhart, Ted. 2000. *Moral Uncertainty and its Consequences*. Oxford: Oxford University Press.

Loewenstein, George, Cindy Bryce, David Hagmann, and Sachin Rajpal. 2015. Warning: You are about to be nudged. *Behavioral Science and Policy* 1 (1): 35–42.

Matthews, Steve. 2015. Bio-Technical Challenges to Moral Autonomy. In *Super Soldiers: The Ethical, Legal and Social Implications*, eds. Jai Galliott and Mianna Lotz, 109–119. Burlington: Ashgate.

McCrudden, Christopher, and King, Jeff. 2016. The Dark Side of Nudging: The Ethics, Political Economy, and Law of Libertarian Paternalism. In *Choice Architecture in Democracies: Exploring the Legitimacy of Nudging*, eds. Alexandra Kemmerer, Christoph Möllers, Maximilian Steinbeis, and Gerhard Wagner, 75–139. Baden-Baden: Nomos Verlagsgesellschaft.

McNeil, Barbara J., Stephen G. Pauker, Harold C. Sox, and Amos Tversky. 1982. On the Elicitation of Preferences for Alternative Therapies. *The New England Journal of Medicine* 306 (21): 1259–1262.

McPherson, Michael S., and Matthew A. Smith. 2008. Nudging for Liberty: Values in Libertarian Paternalism. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1220323.

Mele, Alfred R. 1995. *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford University Press.

Mercier, Hugo, and Dan Sperber. 2009. Intuitive and Reflective Inferences. In *In Two Minds: Dual Processes and Beyond*, eds. Jonathan St.B.T. Evans, and Keith Frankish, 149–170. Oxford: Oxford University Press.

Mill, John Stuart. 1859 [1991]. On Liberty. In *On Liberty and Other Essays* (World's Classics), ed. John Gray, 5–131. Oxford: Oxford University Press.

——. 1861 [1991]. Considerations on Representative Government. In *On Liberty and Other Essays* (World's Classics), ed. John Gray, 203–467. Oxford: Oxford University Press.

Miller, David. 2011. Taking Up the Slack? Responsibility and Justice in Situations of Partial Compliance. In *Responsibility and Distributive Justice*, eds. Carl Knight and Zofia Stemplowska, 230–245. Oxford: Oxford University Press.

Mills, Chris. 2015. The Heteronomy of Choice Architecture. *Review of Philosophy and Psychology* 6 (3): 495–509.

——. 2018. The Choice Architect's Trilemma. *Res Publica* 24 (3): 395–414.

Mitchell, Gregory. 2005. Libertarian Paternalism is an Oxymoron. *Northwestern University Law Review* 99 (3): 1245–1278.

Moles, Andrés. 2015. Nudging for Liberals. *Social Theory and Practice* 41 (4): 644–667.

Mullainathan, Sendhil, and Eldar Shafir. 2013. *Scarcity: Why Having Too Little Means So Much*. New York: Times Books, Henry Holt and Company.

Myers, David. 2002. *Intuition: Its Powers and Perils*. New Haven: Yale University Press.

Nagatsu, Michiru. 2015. Social Nudges: Their Mechanisms and Justification. *Review of Philosophy and Psychology* 6 (3): 481–494.

Nolan, Jessica M., P. Wesley Schultz, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. 2008. Normative Social Influence is Underdetected. *Personality and Social Psychology Bulletin* 34 (7): 913-923.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. Oxford: Blackwell Publishers Ltd.

Nys, Thomas, and Bart Engelen. 2017. Judging Nudging: Answering the Manipulation Objection. *Political Studies* 65 (1): 199–214.

Oliver, Adam. 2013. From Nudging to Budging: Using Behavioural Economics to Inform Public Sector Policy. *Journal of Social Policy* 42 (4): 685–700.

Paulo, Norbert, and Jan Christoph Bublitz. 2019. How (not) to Argue For Moral Enhancement: Reflections on a Decade of Debate. *Topoi* 38 (1): 95–109.

Persson, Ingmar, and Julian Savulescu. 2012. *Unfit for the Future: The Need for Moral Enhancement*. Oxford: Oxford University Press.

——. 2015. The Art of Misunderstanding Moral Bioenhancement: Two Cases. *Cambridge Quarterly of Healthcare Ethics* 24 (1): 48–57.

——. 2016a. Moral Bioenhancement, Freedom and Reason. *Neuroethics* 9 (3): 263–268.

——. 2016b. Enharrisment: a Reply to John Harris about Moral Enhancement. *Neuroethics* 9 (3): 275–277.

——. 2019. The Duty to be Morally Enhanced. *Topoi* 38 (1): 7–14.

Petrescu, Dragos C., Gareth J. Hollands, Dominique-Laurent Couturier, Yin-Lam Ng, Theresa M. Marteau. 2016. Public Acceptability in the UK and USA of Nudging to Reduce Obesity: The Example of Reducing Sugar-Sweetened Beverages Consumption. *PLoS ONE* 11 (6): e0155995. Available at: doi.org/10.1371/journal.pone.0155995.

Pettit, Philip. 1996. Freedom as Antipower. *Ethics* 106 (3): 576–604.

Pichert, Daniel, and Konstantinos V. Katsikopoulos. 2008. Green Defaults: Information Presentation and Pro-environmental Behaviour. *Journal of Environmental Psychology* 28 (1): 63–73.

Planet Money. 2014. Episode 590: The Planet Money Workout. www.npr.org (December 17, 2014). Available at: https://www.npr.org/sections/money/2014/12/17/371463435/episode-590-the-planet-money-workout.

Ploug, Thomas, and Søren Holm. 2015. Doctors, Patients, and Nudging in the Clinical Context – Four Views on Nudging and Informed Consent. *The American Journal of Bioethics* 15 (10): 28–38.

Pogge, Thomas W. 2005. Severe Poverty as a Violation of Negative Duties. *Ethics and International Affairs* 19 (1): 55–83.

——. 2007. *World Poverty and Human Rights*. 2nd edition. Cambridge: Polity.

Posner, Richard A. 2013. Why Is There No Milton Friedman Today? *Econ Journal Watch* 10 (2): 210–213.

Pronin, Emily, Jonah A. Berger, and Sarah Molouki. 2007. Alone in a Crowd of Sheep: Asymmetric Perceptions of Conformity and Their Roots in an Introspection Illusion. *Journal of Personality and Social Psychology* 92 (4): 585–595.

Pugh, Jonathan. 2017. Moral Bio-enhancement, Freedom, Value and the Parity Principle. *Topoi* 38 (1): 73–86.

Quong, Jonathan. 2010. *Liberalism Without Perfection*. New York: Oxford University Press.

Rachlinski, Jeffrey J. 2017. How I Learned to Stop Worrying and Love Nudges (Book Review of *The Ethics of Influence*: *Government in the Age of Behavioral Science*, By Cass Sunstein. New York, New York: Cambridge University Press, 2016). *Texas Law Review* 95 (5): 1061–1076.

Radoilska, Lubomira. 2012. Autonomy and Ulysses Arrangements. In *Autonomy and Mental Disorder*, ed. Lubomira Radoilska, 252–280. Oxford: Oxford University Press.

Rawls, John. 1996. *Political Liberalism*. New York: Columbia University Press.

——. 1999a. *A Theory of Justice, Revised Edition*. Oxford: Oxford University Press.

——. 1999b. *The Law of Peoples*. Cambridge: Harvard University Press.

——. 2001. *Justice as Fairness: A Restatement*, ed. Erin Kelly. Cambridge: Harvard University Press.

Raz, Joseph. 1999. *Practical Reasons and Norms, 2nd edition*. Oxford: Oxford University Press.

Rebonato, Riccardo. 2014. A Critical Assessment of Libertarian Paternalism. *Journal of Consumer Policy* 37 (3): 357–396.

Reisch, Lucia A., and Cass R. Sunstein. 2016. Do Europeans like nudges? *Judgment and Decision Making* 11 (4): 310–325.

Riley, Evan. 2017. The Beneficent Nudge Program and Epistemic Injustice. *Ethical Theory and Moral Practice* 20 (3): 597–616.

Rosati, Connie S. 2016. Moral Motivation. *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Available at: https://plato.stanford.edu/entries/moral-motivation/.

Rosenberg, Matthew, Nicholas Confessore, and Carole Cadwalladr. 2018. How Trump Consultants Exploited the Facebook Data of Millions. *www.nytimes.com* (March 17, 2018). Available at: https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html.

Rothman, Daniel H. 2017. Thresholds of catastrophe in the Earth system. *Science Advances* 3 (9).

Rubin, Zick, and Letitia A. Peplau. 1975. Who Believes in a Just World? *Journal of Social Issues* 31 (3): 65–89.

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

Samuelson, William, and Richard Zeckhauser. 1988. Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty* 1 (1): 7–59.

Sandel, Michael J. 1998. *Liberalism and the Limits of Justice, 2nd edition*. New York: Cambridge University Press.

Savulescu, Julian, and Ingmar Persson. 2012. Moral Enhancement, Freedom, and the God Machine. *The Monist* 95 (3): 399–421.

Scanlon, Thomas M. 1988. The Significance of Choice. In *The Tanner Lectures on Human Values (Volume 8)*, ed. Sterling M. McMurrin, 149–216. Salt Lake City: University of Utah Press.

Schaefer, G. Owen. 2015. Direct vs. Indirect Moral Enhancement. *Kennedy Institute of Ethics Journal* 25 (3): 261–289.

——, Guy Kahane, and Julian Savulescu. 2014. Autonomy and Enhancement. *Neuroethics* 7 (2): 123–136.

Schmidt, Andreas T. 2017. The Power to Nudge. *American Political Science Review* 111 (2): 404–417.

Schubert, Christian. 2015. On the ethics of public nudging: Autonomy and agency. *Joint Discussion Paper Series in Economics*, 33-2015.

——. 2016. Green nudges: Do they work? Are they ethical? *Joint Discussion Paper Series in Economics*, 09-2016.

Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. New York: Academic Press.

Selinger, Evan, and Kyle P. Whyte. 2012. Nudging Cannot Solve Complex Policy Problems. *European Journal of Risk Regulation* 3 (1): 26–31.

Sen, Amartya. 1985. The Moral Standing of the Market. *Social Philosophy and Policy* 2 (2): 1–19.

Sepielli, Andrew. 2013. Moral Uncertainty and the Principle of Equity among Moral Theories. *Philosophy and Phenomenological Research* 86 (3): 580–589.

Shell, Ellen R. 2009. *Cheap: The High Cost of Discount Culture*. New York: The Penguin Press.

Simmons, A. John. 2010. Ideal and Nonideal Theory. *Philosophy and Public Affairs* 38 (1): 5–36.

Singer, Peter. 2004. *One World: The Ethics of Globalization, 2nd edition*. New Haven: Yale University Press.

Slovic, Paul. 2010. If I Look at the Mass I Will Never Act: Psychic Numbing and Genocide. In *Emotions and Risky Technologies*, ed. Sabine Roeser, 37–59. The International Library of Ethics, Law and Technology 5. Dordrecht: Springer.

Small, Deborah A., and George Loewenstein. 2003. Helping *a* Victim or Helping *the* Victim: Altruism and Identifiability. *Journal of Risk and Uncertainty* 26 (1): 5–16.

——, and Paul Slovic. 2007. Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes* 102 (2): 143–153.

Sparrow, Robert. 2014. Better living through chemistry? A reply to Savulescu and Persson on "moral enhancement". *Journal of Applied Philosophy* 31 (1): 23–32.

Stanovich, Keith E. 2004. *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. Chicago: Chicago University Press.

——. 2009. Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In *In Two Minds: Dual Processes and Beyond*, eds. Jonathan St.B.T. Evans, and Keith Frankish, 55–88. Oxford: Oxford University Press.

Stemplowska, Zofia, and Adam Swift. 2012. Ideal and Nonideal Theory. In *The Oxford Handbook of Political Philosophy*, ed. David Estlund, 373–390. Oxford: Oxford University Press.

Strahan, Erin J., Katherine White, Geoffrey T. Fong, Leandre R. Fabrigar, Mark P. Zanna, and Roy Cameron. 2002. Enhancing the effectiveness of tobacco package warning labels: a social psychological perspective. *Tobacco Control* 11 (3): 183–190.

Sugden, Robert. 2018. *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford: Oxford University Press.

Sunstein, Cass R. 2006. Boundedly Rational Borrowing. *University of Chicago Law Review* 73: 249–270.

——. 2013. *Simpler: The Future of Government*. New York: Simon & Schuster.

——. 2015a. Nudges, Agency, and Abstraction: A Reply to Critics. *Review of Philosophy and Psychology* 6 (3): 511–529.

——. 2015b. *Why Nudge? The Politics of Libertarian Paternalism*. New Haven & London: Yale University Press.

——. 2016a. The Ethics of Choice Architecture. In *Choice Architecture in Democracies: Exploring the Legitimacy of Nudging*, eds. Alexandra Kemmerer, Christoph Möllers, Maximilian Steinbeis, and Gerhard Wagner, 21–74. Baden-Baden: Nomos Verlagsgesellschaft.

——. 2016b. *The Ethics of Influence: Government in the Age of Behavioral Science*. Cambridge: Cambridge University Press.

——. 2016c. Do People Like Nudges? *Administrative Law Review* 68 (2): 177–232.

——. 2017. *Human Agency and Behavioral Economics: Nudging Fast and Slow*. Cham: Palgrave Macmillan.

——. 2019. *On Freedom*. Princeton, Oxford: Princeton University Press.

Sunstein, Cass R., and Edna Ullmann-Margalit. 1999. Second-Order Decisions. *Ethics* 110 (1): 5–31.

——, and Lucia A. Reisch. 2013. Green by Default. *Kyklos* 66 (3): 398–402.

——, Lucia A. Reisch, and Julius Rauber. 2018. A worldwide consensus on nudging? Not quite, but almost. *Regulation and Governance* 12 (1): 3–22.

Svavarsdottir, Sigrún. 1999. Moral Cognitivism and Motivation. *The Philosophical Review* 108 (2): 161–219.

Tadros, Victor. 2011. *The Ends of Harm: The Moral Foundations of Criminal Law*. Oxford, New York: Oxford University Press.

Thaler, Richard H. 1980. Toward a Positive Theory of Consumer Choice. *Journal of Economic Behavior and Organization* 1 (1): 39–60.

——. 1991. *The Winner's Curse: Paradoxes and Anomalies of Economic Life*. New York: Free Press.

——, and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven: Yale University Press.

Thompson, Suzanne C. 1999. Illusions of Control: How We Overestimate Our Personal Influence. *Current Directions in Psychological Science* 8 (6): 187–190.

Thomson, Judith J. 1985. The Trolley Problem. *The Yale Law Journal* 94 (6): 1395–1415.

Tomasi, John. 2012. *Free Market Fairness*. Princeton, Oxford: Princeton University Press.

Trout, J.D. 2009. A restriction maybe, but is it paternalism? Cognitive bias and choosing governmental decision aids. *NYU Journal of Law & Liberty* 2-3: 455–469.

Tsai, George. 2014. Rational Persuasion as Paternalism. *Philosophy and Public Affairs* 42 (1): 78–112.

Tversky, Amos, and Daniel Kahneman. 1974. Judgment Under Uncertainty: Heuristics and Biases. *Science* 185 (4157): 1124–1131.

Ubel, Peter A., Dylan M. Smith, Brian J. Zikmund-Fisher, Holly A. Derry, Jennifer McClure, Azadeh Stark, Cheryl Wiese, Sarah M. Greene, Aleksandra Jankovic, and Angela Fagerlin. 2010. Testing Whether Decision Aids Introduce Cognitive Biases: Results of a Randomized Trial. *Patient Education and Counseling* 80 (2): 158–163.

Valdman, Mikhail. 2010. Outsourcing Self-Government. *Ethics* 120 (4): 761–790.

Valentini, Laura. 2012. Ideal vs. Non-Ideal Theory: A Conceptual Map. *Philosophy Compass* 7 (9): 654–664.

Waldron, Jeremy. 1981. A Right to Do Wrong. *Ethics* 92 (1): 21–39.

——. 2014. It's All for Your Own Good. *The New York Review of Books*. *www.nybooks.com* (October 9, 2014). Available at: https://www.nybooks.com/articles/2014/10/09/cass-sunstein-its-all-your-own-good/.

Wall, Stephen. 2016. Autonomy as a Perfection. *American Journal of Jurisprudence* 61 (2): 175–194.

Wedel, Michel, and Rik Pieters. 2008. A Review of Eye-Tracking Research in Marketing. *Review of Marketing Research* 4: 123–147.

Wegner, Daniel M. 2002. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Wendel, Steve. 2016. Behavioral Nudges and Consumer Technology. In *Nudge Theory in Action: Behavioral Design in Policy and Markets*, ed. Sherzod Abdukadirov, 95–123. Palgrave Macmillan.

White, Mark D. 2013. *The Manipulation of Choice: Ethics and Libertarian Paternalism*. New York: Palgrave Macmillan.

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

World Health Organization. 2018. *Global status report on road safety 2018*. Geneva: World Health Organization.

Yudkowsky, Eliezer. 2008. Cognitive Biases Potentially Affecting Judgment of Global Risks. In *Global Catastrophic Risks*, eds. Nick Bostrom and Milan M. Ćirković, 91–119. New York: Oxford University Press.

Yeung, Karen. 2012. Nudge as Fudge. *The Modern Law Review* 75 (1): 122–148.

Young, Robert. 1980. Autonomy and the 'Inner Self'. *American Philosophical Quarterly* 17 (1): 35–43.

Zajonc, Robert B., and D.W. Rajecki. 1969. Exposure and affect: A field experiment. *Psychonomic Science* 17 (4): 216–217.

Zohny, Hazem. 2018. Moral enhancement and the good life. *Medicine, Health Care and Philosophy* 22 (2): 267–274.