Living Well by Design:
An Account of Permissible Public Nudging

Fay Niker

Thesis submitted for the degree of Doctor of Philosophy
Politics and International Studies
University of Warwick
June 2017

Contents

Acknowledgements

First and foremost, I would like to thank my supervisors – Matthew Clayton and Adam Swift – for their intellectual guidance and support during this project. This thesis owes an immeasurable amount to their patient and conscientious engagement with my work. Adam has now supervised both my M.Phil and PhD theses, and any confidence I have gained in my work over these six years can probably be attributed in some way to his invaluable support and reassurance. I am immensely grateful to Matthew for being such a generous supervisor; he is always available, and never makes you feel like he has more important things to be doing (even when he does!). Simply put, I could not have wished for better supervisors.

There are numerous other people, both at Warwick and further afield, who have made my PhD years a fun and fruitful experience. First, Warwick's Centre for Law, Ethics and Public Affairs (CELPA) has provided both an academically stimulating environment and an important sense of community. Second, I wish to thank Kimberley Brownlee, for whom I worked as a research assistant for part of my degree. I very much enjoyed this varied work, and it was always a pleasure to work with Kim. Both Kim and Zosia Stemplowska have been, and continue to be, role models of mine. Third, I have had the chance to work on some collaborative projects during my PhD, and have learned vast amounts from my co-authors, Peter Reiner and Gidon Felsen. Particular thanks go to Peter for responding enthusiastically to my initial email and for hosting me at the National Core for Neuroethics during my visit to Vancouver. Fourth, I have been very fortunate to have made many friends through my doctoral studies, including: Cristina Astier, Christopher Bennett, Joseph Burke, Nick Fitz, Caroline Harnacke, Thomas Hartvigson, Clare Heyward, Helen McCabe, Viktor Ivanković, Julia Mosquera, Tom Parr, Enric Bea Seguí, Priya Sury, Isa Trifan, and Kelly Tuke. I am also indebted to Viktor for being willing to proofread the thesis, and for doing such a thorough job.

Pursuing this PhD has only been possible due to the support of my family. I owe the most to my mother, who has been a constant source of love and support. One often finds themselves reaping the rewards from seeds that others have sown. This PhD is one example of this, and I thank my mother, Mary Niker, and my grandparents, Ann and James Towey, in particular for sowing so many seeds into my life.

Without doubt the best thing to happen during my time at Warwick was meeting my partner, David. He both inspires and grounds me, and has supported me incalculably, especially in these final few months.

Finally, this research has been funded by the Economic and Social Research Council (ESRC) and by a Chancellor's Scholarship from the University of Warwick, as well as by a grant from the Horowitz Foundation for Social Policy. I am very grateful to these three institutions for this financial support; this project would not have been possible without it. I have also been very fortunate to spend time at the University of British Columbia and the Universitat Pompeu Fabre in Barcelona, both of which were exceedingly fruitful exchanges. I thank the Humanities Research Centre for the Warwick Transatlantic Fellowship that funded the former, and the ESRC for the Overseas Institutional Visit grant that supported the latter.

Declaration

I declare that this thesis is my own work and that it has not been submitted for a degree at another university. It does not contain work published elsewhere.

Abstract

The thesis provides a full assessment of the moral permissibility of a set of new belief and behaviour modification techniques, now commonly known as "nudges", which are grounded in and justified by reference to our new insights into human psychology. It asks what forms of nudging are permissible in light of the state's new understanding of its capacity to modify behaviour using these insights; and it develops an ethico-political account of living well that directs this normative investigation. There are two main strands to this analysis of public nudging, one relating to behaviour change policies designed for the sake of the target and the other relating to those designed for the sake of others. Across both strands, it is argued that the kinds of interventions that are permissible share a similar character: specifically, they are compatible with creating and sustaining the conditions for living well, on account of their playing an ecological-educative role in supporting citizens' personal autonomy and practical reasoning. The thesis uses its in-depth normative analysis as the basis for engaging with current practices in behavioural policymaking and for setting out an ethically-sensitive policy framework to guide the design of nudge interventions in practice. The extended argument presented in these pages offers a distinctive and timely contribution to this debate, setting out arguably the most sustained and complete philosophical assessment of the ethics of nudging in the literature to date.

1. Introduction: The Ethics of Public Influence

"Love it, or hate it, nudging is here to stay." (Halpern 2015a: 12)

Our understanding of human agency has changed dramatically in recent years as the result of a robust set of discoveries in the cognitive and behavioural sciences. These discoveries, concerning the relationship between our conscious and subconscious selves and the complex interface between the rational and irrational (Whitehead et al. 2011: 2819), have been at the heart of "a quiet revolution taking place in government" (Sanders and Halpern 2014). Knowledge of the variety of cognitive biases affecting our thinking and the deep impact of seemingly trivial situational factors on our behaviour has opened up new conceptual space within which to explore the possibility of innovative policy tools that work "with the grain" of our psychology (Cabinet Office 2010: 2).[1] The most influential of these rests on the idea of designing "choice architecture" in psychologically-informed ways in order to "nudge" people into making certain welfare-promoting choices (Thaler and Sunstein 2009). As a simple matter of fact, this application of cognitive and behavioural science to public policy has been reshaping the operational logics of policymaking in governments around the world – a shift referred to within policy circles as the *behaviour change agenda*, but which has been described, more critically, as "the rise of the psychological state" (Jones et al. 2013).

This development has been grounded in and putatively justified by a philosophical argument claiming that this nudge policy programme should be deemed an acceptable, even desirable, form of paternalism because the policies it recommends do not offend against liberal values (Thaler and Sunstein 2009; Sunstein 2014a, 2016a). This argument has played a part in reopening central debates in liberal political theory over what should count as permissible governmental interference with citizens' behaviour. John Stuart Mill refers to this as the "practical question" of "how to make the fitting adjustment between individual independence and social control" (2005: 8). He opens *On Liberty* by stating that this is "a subject on which nearly everything remains to be done" (ibid.); his analysis in the pages that follow, however, laid the foundations for liberalism's anti-paternalist orthodoxy, which has remained largely unchallenged

[1] This psychological evidence and the shift to thinking about these findings as "behavioural insights" for policymaking is outlined in full in §2.1.

until recent years. Has the naturalistic challenge to our prevailing conceptions of human agency called into question, or even undermined, this orthodox view on the ethics of public influence? Given these empirical findings, how should we think about the moral permissibility of designing policy environments based on knowledge of our predictable irrationalities for the purpose of improving citizens' lives?

These are complex questions about which much "remains to be done". What is more, there is some urgency about the need to find philosophically coherent answers to these questions, particularly given the daily extension of the behavioural policy programme and the potential normative costs that are attached to this development. To this end, the main aim of this thesis is to defend an attractive and complete account of the conditions under which it is permissible for governments to seek to modify their citizens' behaviour via nudge policies. In so doing, it will make important progress in clarifying the normative dimensions of this policy development and in presenting a comprehensive and philosophically sophisticated assessment of these issues.

This introductory chapter has three main aims. The first is to highlight the political and philosophical timeliness of this type of investigation, which serves further to motivate the need for the inquiry carried out by this thesis (§1.1.1 and §1.2.1). The second is to lay the foundations for the remainder of the thesis by issuing some remarks relating to the scope of this inquiry and the methods it uses, which serve to clarify the thesis's main assumptions (§1.1.2 and §1.2.2). The third is to provide a brief preview of each of the chapters and their main conclusions, which affords me the opportunity to sketch the shape of the extended argument and to clarify the links between chapters (§1.3).

§1.1: Changing Citizens' Behaviour

Governments have always sought to influence the behaviour of those they govern, and have numerous tools with which to affect this influence, ranging from strong forms of coercion to educative means such as information provision and rational persuasion. In recent years, however, a distinctive approach to public influence has begun to command increasing interest from policymakers and social planners, and has had a significant impact on policymaking. This section first offers an overview of this "behavioural revolution" (Leggett 2014: 4), in order to illustrate its extensive reach and momentum as a policy logic (§1.1.1). Given this extensive reach, the second part of this section offer

some comments on the specific type of influencing activities that are the focus of this normative inquiry, which serves to clarify the scope of the thesis (§1.1.2).

§1.1.1: Nudge: A New and Growing Movement

By far the most influential account of behavioural policy to date is that advanced by Richard Thaler and Cass Sunstein in their book *Nudge*, which, to their minds, offers a manual for governments wanting to work with the grain of how people actually think and behave with the aim of making designing more effective policy. They have labelled their "new movement" *libertarian paternalism*, because it seeks to bring together the seemingly oxymoronic claims that governments should respect people's freedom to choose and that it is legitimate for governmental actors to try to steer people's choices in ways that will improve their lives (2009: 5). Its core tenet is that, in response to the knowledge that people are "so susceptible to various counterproductive behavioural and cognitive tendencies" (Anderson 2010: 369), it is appropriate for policy makers to use behavioural insights to modify decision-making situations – or, in their terms, to design "choice architecture" – in ways that steer citizens' behaviour "in the right direction (as judged by people themselves)" (Sunstein 2014a: 17). Interventions of this kind are labelled "nudges", and are defined as "any aspect of the choice architecture that alters people's behaviour in predictable ways without forbidding any options or significantly changing their […] incentives" (Thaler and Sunstein 2009: 6).[2] The argument in favour of libertarian paternalism has been enthusiastically accepted by policymakers, making nudging "one of the hottest ideas in current policy debates" (Hausman and Welch 2010: 123).

On coming to power in 2010, for instance, former Prime Minister of the United Kingdom (UK) David Cameron lost little time in establishing the Behavioural Insights Team (BIT) – more commonly known as the "nudge unit" – within the Cabinet Office, its mandate being "to help the UK Government develop and apply lessons from behavioural economics and behavioural science in policymaking" (O'Donnell 2010). Since then, the logic of behavioural insights has been built into the institutional framework and bureaucratic mind-sets of "almost every government department" (ibid.), with Richard Thaler acting as a principal adviser for this process. Indeed, the government's "preference for non-regulatory interventions" was quite quickly deemed

[2] A full account of the argument in favour of this form of paternalism is presented, and engaged with critically, in the next chapter (§2.2).

strong enough to warrant a House of Lords Select Committee investigation focused on assessing the evidence base for this shift, as well as examining some aspects relating to the ethics of the techniques used in behaviour change policy (House of Lords 2011). The BIT was semi-privatised in early 2014, taking on the form of a mutual joint venture or "social purpose company" owned in roughly equal proportions by its employees, the UK government, and Nesta (the government-established "innovation charity" where it is now housed).[3] The privatisation deal was described by *The Financial Times* as "one of the biggest experiments in British public sector reform", since this is the first time that privatisation has reached beyond public services and utilities to include an actual government policy team (Plimmer 2014).[4] The result has been that the BIT acts not only as a "*de facto* Council of Psychological Advisers" to the UK government (Halpern 2015b: 768), but also in the capacity of a global behavioural policy consultancy firm for foreign governments, private sector clients, and other institutions. Within domestic policy, the BIT has overseen a number of changes across a wide range of areas, including health care, tax and fraud, personal savings, charitable giving, energy consumption and sustainability, and crime reduction (Cabinet Office 2010, 2012, 2013; BIT 2015, 2016). It is plausible to think that, unless something significant occurs, this influence is set only to grow further. Indeed, as two senior BIT members write, "Although the application of behavioural sciences to public policy is in its relative infancy, there are signs that this may prove an enduring development" (Hallsworth and Sanders 2016: 126).

The approach is also beginning to wield a similar level of influence in the United States (US). During President Obama's first term in office, Cass Sunstein served as the Administrator of the White House Office of Information and Regulatory Affairs (OIRA), earning him the title "regulation czar".[5] In late 2015, Obama issued an Executive Order entitled "Using Behavioral Science Insights to Better Serve the

[3] Nesta is an independent foundation, which was established by the last Labour government using £250 million of National Lottery money. Incidentally, its Chief Executive Officer is Geoff Mulgan, who worked closely with David Halpern, the BIT's CEO, in the UK government's Strategy Unit when Tony Blair was Prime Minister. For more information, see: http://www.nesta.org.uk.

[4] The privatisation was due, at least in part, to a cap on the number of BIT employees and to its employees pay being bound to civil service pay grades. It also enabled the BIT to cash in on its competitive advantage and set itself up as, in effect, an international consultancy company charging its clients "success fees" (i.e., if the advice it provides leads to significant client savings, the company gets a share).

[5] Sunstein's time in office was by no means uncontroversial, however. According to Rena Steinzor, president of the Center for Progressive reform and professor at the University of Maryland's Law School, "Cass Sunstein is the most well-connected and smartest guy who's ever held the job. But he's also done untold damage" (quoted in Broder 2012).

American People" (#13707), which directed all federal agencies to incorporate this evidence into their policymaking logics and processes. In conjunction with this, the US formally established the Social and Behavioral Sciences Team (SBST), dedicated to supporting the various agencies and departments in crafting policies that are "built for people" in light all of their psychological foibles.

Similar trends have taken place in many other countries, including Germany, Australia, Denmark, Singapore, the Netherlands, and Israel, in addition to international organisations. The World Bank, for example, devoted its 2015 World Development Report (entitled *Mind, Society, and Behaviour*) to discussing the implications of behavioural economics and policy for international development, with an eye to making progress on "a new set of development approaches based on a fuller consideration of psychological and social influences" (World Bank 2015: 2). Similarly, the United Nations Development Programme recently established the UN Behavioural Initiative, charged with "translating behavioural science insights into more effective and efficient UN programming and operations", in the hope that this will have a positive impact on its efforts to reach the Agenda 2030 sustainable development goals (UN 2016: 3). The OECD, too, published a report only in the last few months – *Behavioural Insights and Public Policy: Lessons from Around the World* – that presents the results of "a first-of-its-kind survey" of behavioural insights units (OECD 2017). This reach has expanded further still: Maya Shankar, the founder of the SBST and the first-ever Behavioural Science Advisor to the UN (a role she performed alongside her White House position), became Google's first Head of Behavioral Insights in February 2017, which marks the next expansion of behavioural policy.

It is clear, then, that nudging is "a growing movement" (Sunstein 2016a: 5). But what ethical principles, if any, guide and constrain this onward march of the behavioural policy agenda? Has there been any serious assessment by nudge advocates of the potential normative costs attached to the use of these forms of influence? Have clear distinctions been drawn relating to the different ways in which behavioural policies might act on their targets, and what this means for their permissible use? Unfortunately the answers to these questions reveal a rather worrying gap in current thinking on behavioural policymaking. Engagement with these ethical dimensions has often tended to be cursory at best, and follows a distinct trend: Thaler and Sunstein offer a "breezy discussion of objections" (Anderson 2010: 371), mainly pragmatic rather than principled, at the end of the *Nudge* (2009: 235-252); the World Bank's report includes a

very brief discussion on the question of 'Why should governments shape individual choices?', more as an afterthought in the final couple of pages it would seem (2015: 202-203); and even the House of Lords Select Committee relegated its discussion of ethical issues to the last of its six Appendices (2011: 105-109).

The principal advocates of nudging have only very recently joined the debate over its potential normative costs (Schubert 2015: 3) – something that it seems they were keen to avoid, but which has become necessary in light of sustained criticisms presented both in the academic literature (e.g., Hausman and Welch 2010; Rebonato 2012; Grüne-Yanoff 2012; White 2013) and in the media (Waldron 2015; Shaw 2017). There are a plethora of different worries that have been raised; but the two leading objections are that the nudge policy programme is a disrespectful form of paternalism and thereby an affront to human dignity, and that nudges act on people in a manipulative fashion, which undermines their personal autonomy. These concerns are captured in extracts from two pieces written for *The New York Review of Books*:

> "Nudging doesn't teach me not to use inappropriate heuristics or to abandon irrational intuitions… I still use the same defective strategies but now things have been arranged to make that work out better. Nudging takes advantages of my deficiencies in the way one indulges a child." (Waldron 2015)

> "The behavioral techniques that are being employed […] do not appeal to our reason; they do not seek to persuade us consciously with information and argument. Rather, these techniques change behaviour by appealing to nonrational motivations, our emotional triggers and unconscious biases." (Shaw 2017)

In late 2016 Sunstein published his first sustained attempt to integrate these sorts of normative considerations into his pro-nudge argument. Nevertheless, as reviewers of *The Ethics of Influence* have highlighted, while the book offers "an excellent primer of the relevant debates on nudges", it fails to provide "a bona fide philosophical inquiry into the underlying ethical issues", which is precisely "the sort of investigation that is sorely needed" (Johnson 2016). This thesis aims to offer this "sorely needed" comprehensive normative investigation into this politically pressing issue.

§1.1.2: Scope of Inquiry

Before setting out my approach to this principled evaluation, which I do in the next section, it is instructive to offer some initial remarks about the scope of the inquiry

carried out in this thesis. Two main points are worth noting at this stage, for the sake of clarity going forward.

First, we saw above how expansive the reach of behavioural policy has become. My inquiry focuses on the use of nudges at the level of national public policymaking. It explores the conditions under which it is permissible for a state to influence its citizens' beliefs or behaviour in this way.[6] Throughout the thesis, I will often refer to this as *public nudging* or *public choice architecture*, and it is worth making clear at the outset what these mean so as to avert any potential confusion.[7] The prefix "public" denotes that the nudge is a governmental intervention, designed and delivered by public authorities; it does not cover all forms of nudging done in public places. Public authorities can be involved in nudging citizens in more and less direct ways. They might design policies that influence people directly, such as building prompted choice for organ donation registration into the application for vehicle tax. But a government might also intervene in a more indirect way vis-à-vis individual citizens. It may direct its actions toward other public actors with the aim of guiding the choices of individual citizens, such as when it introduces corporate regulation relating to the salient presentation of nutritional information on food items or public health warnings on cigarette packets. Both fit with the definition of public nudging given above. A final clarificatory point on "public choice architecture": since "public choice" has its own associations, this phrase may be ambiguous between two possible locutions, "the architecture of public choice" and "choice architecture that is public"; I hope that my stipulation above has made it clear that the second of these is what interests us here.

Behavioural policy has proven attractive to policymakers partly because it offers a relatively cheap course of action, both financially[8] and politically, at a time when (i) government budgets are under pressure, and (ii) it can be difficult to pass stronger regulatory measures on account of the logjams resulting from highly politicised divides between left- and right-wing positions. These political contextual considerations help us

[6] Other actors use these techniques of influence. Nudge-style influences play a key role, for instance, in the marketing and advertising activity of many private corporations. Such influences also need to be subjected to serious normative attention – which is something that I discuss in more detail in §6.4. There are two principled reasons for focusing on the influencing actions of states or political authorities in the first instance: the first relates to the general authority that the government of a state claims for itself, including the rules governing the activities of other corporations; the second is that a government is entitled only to pursue the interest of its subjects, as opposed to private corporations which are entitled to pursue, to some extent at least, their own independent interests (Raz 1986: 4-6).

[7] It may be that nudging is only one type of choice architecture. That is, the latter could be understood as covering a wider range of possible influences. Here, however, I use the two terms as interchangeable.

[8] Halpern and Sanders claim that the BIT "is delivering monetary benefits in the region of hundreds of millions of dollars, if not billions" (2016: 54).

to understand why the nudge policy programme has been so enthusiastically adopted at this particular point in time, and why this has been so in the case of some governments more than others.[9] Understanding and critically engaging with these conditions as they relate to debates on the nature of contemporary governance is a valuable inquiry; but it is the not the type of inquiry carried out in this thesis.[10] A second consideration about scope, then, concerns the methods I adopt. As is explained in more detail below, my inquiry is focused on examining the *normative underpinnings* of this adoption by governments of new behaviour change strategies, with the aim of presenting an account of morally permissible nudging.

§1.2: A Principled Evaluation

I hope so far to have highlighted the political timeliness of this inquiry. It is not only that nudging appears to be "here to stay", as the BIT's Chief Executive David Halpern confidently asserts in this chapter's opening quotation; he is also quoted as claiming that he and his team of policy entrepreneurs "have a sharp sense that we are still just scratching the surface" and that there is "so much more we can do" (quoted in Rutter 2015). I agree that it is difficult to deny that there is a role, perhaps even an important one, for public choice architecture within an overall approach to policymaking; nonetheless, in order to understand what this role should be, we need an account of the political principles that ought to guide behavioural policymaking. And, to put it simply, "much work still needs to be done in identifying [these] principles" (Anderson 2010: 376). This section offers two sets of remarks about the approach I take to this task. The first is methodological; it details the nature of the normative investigation that I take to be necessary and, in doing so, shows how distinctive this approach is to the one taken by nudge advocates, especially Sunstein (§1.2.1). The second is more substantive; it sets out some foundational claims concerning the demands of political morality, thereby clarifying the theoretical premises on which my extended argument rests (§1.2.2).

§1.2.1: In Praise of Theory

The question at the heart of the thesis's investigation, expressed in its general form, is: Under what conditions, if any, is it morally permissible for a government to use insights from the cognitive and behavioural sciences to design policies with the express aim of

[9] For a very interesting, first-hand story about how the BIT came into existence only after earlier political hiccups, see Halpern (2015a: 301-303).
[10] For examples of this more sociological approach, see Brown (2012) and Leggett (2014).

nudging the behaviour of its citizens in particular directions? Most broadly defined, this is a question that belongs to the discipline of practical philosophy, although this category can be subdivided into three distinct branches according to the sort of practical problems they discuss. The first, value theory, is concerned with investigating the goodness or badness of various actual or possible situations and identifying their good- or bad-making features. The second, normative theory, is concerned primarily with determining what various actors ought to do; it attempts to identify reasons for actions, duties, and rights as a way of establishing the moral permissibility or impermissibility of certain actions. The third, ascriptive theory, is concerned with analysing the conditions for the ascription of blame and praise to various actors (Raz 1999: 11-12). As the phrasing of my question suggests, this thesis is an exercise in *normative theory*, as it is concerned with a specific set of normative questions related to the moral permissibility of public nudging. In seeking to find principles that explain why the use of choice architecture as a policy lever is permissible under certain conditions but not others, the thesis sets out primarily to defend an account of the political morality of nudging.

This methodological approach is starkly juxtaposed to that taken by advocates of nudging in their recent engagement with its normative dimensions. Sunstein repeatedly warns of the need to avoid the "trap of abstraction", which he says can "cause serious confusion when we are thinking about public policy" (2016a: 26). This charge is based on his longstanding commitment to pragmatic, rather than principled, ways of approaching practical questions in law and politics. He was involved in a debate in 1990s, for instance, about the appropriate way to reason about the truth claims of law. This debate pitched "theory-embedded" views, such as that of Ronald Dworkin (1986, 1997), against an "anti-theory army", led by the Chicago School, and with Sunstein (1995, 1996) and Richard Posner (1995) as its leaders (Dworkin 1997: 354, 361). It is helpful briefly to sketch these two views, as this will shed light on the key differences in the approach taken in this thesis, vis-à-vis that taken by nudge advocates, and will offer the opportunity to suggest some reasons for favouring the former over the latter.

The general shape of the pragmatic approach to legal reasoning, and by extension the "practical approach to government" at the base of the pro-nudge argument (Halpern 2015a: 7), is captured in Posner's own characterisation, presented as the following list of adjectives: "practical, instrumental, forward-looking, activist, empirical, sceptical, anti-dogmatic, experimental" (1995: 11; cited in Dworkin 1997:

364). The driving idea is that the central, or perhaps the only, question that needs to be addressed is: How can we make things better? Answering this question requires taking pragmatic and empirical measures to gauge the comparative consequences of different decisions; it does not require abstract philosophical reasoning about complex issues of political morality. What matters, in short, is "what works". By contrast, the "theory-embedded" view contends that legal reasoning is the application of principles, which are embedded in practice, to a particular problem with the aim of showing that this or that principle provides a better justification of some part of legal practice (Dworkin 1997). It denies that there can be any *a priori* limit or constraint put on legal reflection "as a boundary of abstraction such reflection must not cross", because it is the character of the inquiry itself which will, in each case, dictate the level of theory that needs to be explored (ibid.: 371-372). The same can be said of political reasoning. Many practical problems in politics require that we engage in abstract, philosophical theorising – and the ethics of public influence is one such example, as we will see in subsequent chapters. We cannot set an *a priori* constraint on what level of theoretical reasoning is and is not appropriate to resolving this practical problem; rather, it is only in considering it that we will discover which general philosophical or theoretical issues need to be confronted to resolve the issue of when, if at all, it is permissible for governments to modify citizens' behaviour using psychologically-informed techniques.

I believe that this theoretical approach is "not only attractive but inevitable" (Dworkin 1997: 355). Thaler and Sunstein's discussion of the normative dimensions of nudging is, in contrast, marked by an agile pragmatism that is designed to appeal to our common sense intuitions about the "sensible" way of responding politically to the scientific evidence of cognitive biases. However, what this approach lacks is sufficient attention to the principles underlying these cases. As Anderson highlights:

> "It can be an effective rhetorical strategy, for some audiences, to imply that the only opposition to your ideas comes from overly abstract theorists who care more about consistent principles than about effective policy. But Thaler and Sunstein also claim to be introducing a new theoretical underpinning for policy-making, and so the objections cannot be so easily dismissed." (Anderson 2010: 371-372)

It would seem that the pragmatist approach is being used to smuggle in a particular conception of moral theory, namely, a welfarist version of consequentialism. We will explore this welfarism and the problems with it in more detail in later chapters. Here, I simply want to draw attention to the fact that, if the pragmatist approach to assessing

the permissibility of nudging is based on some conception of welfare consequentialism, then its advocates owe us an argument for why this is the most appropriate political principle for this evaluation – or, at the very least, they need to provide some answers relating to the many serious objections that it has been charged with. Such an argument would need to engage in precisely the kind of theorising that Sunstein had been hoping to avoid. In fact, in *The Ethics of Influence* (2016), Sunstein chooses not to take this route, but rather to present a primer of the relevant considerations that has been described as "immediately veering onto safe ground" whenever difficult dilemmas arise (Johnson 2016).

This pragmatist avoidance of issues of principle leaves us with a troubling lack of clarity; in essence, the issue of which nudges can be used under what conditions is left chiefly down to the logic of "what works".[11] Interestingly, Sunstein appears somewhat to embrace the description of his new movement as representing the rise of "the Psychological State" (Jones et al. 2013) – a phrase that conjures up thoughts of living in "nudge-world" (Waldron 2015) or "the Republic of Nudge" (Rachlinski 2017: 1063). In such societies, we might imagine that the landscape in which citizens navigate their lives is extensively curated by *psychocrats*, if you will, based on knowledge of the cognitive and affective bases of human behavioural and decision-making processes, and with the principal aim of steering these processes in directions that promote individual and social welfare.[12] Although Sunstein does acknowledge that "the Psychological State" is "not the best advertising, because it seems a bit alarming", he nonetheless maintains that "the term has the virtue of spotlighting efforts […] to develop sensible, low-cost policies with close reference to how human beings actually think and behave" (Sunstein 2014a: 13). But what on Sunstein's pragmatic approach would stop all of the seemingly sensible individual policy interventions from creating, over time, the "totality of nudging" described above (Rachlinski 2017: 1075)? It is not obvious that, in principle, anything would. Given the current pace of the expansion of behavioural policy, it is clear that there is a real need to subject this issue to comprehensive and sophisticated philosophical analysis, in order to find ethical principles to guide the current practice and future development of nudging. This thesis contributes to this effort.

[11] The reference to "which nudges" here hints at an important aspect of my own analysis, namely, that nudges may be very distinct in their form and impact, and that this may play a role in determining how we should approach them from a normative perspective.

[12] Rachlinski gives an excellent and elaborate illustration of what this might look like in practice in the first few pages of his recent extended review of *The Ethics of Influence* (2017: 1061-1063).

The dystopian image evoked above could give an overly negative impression of the type of analysis that this thesis is set to offer. I present it, however, to highlight the problems attendant on an anti-theoretical approach, rather than as an implicit critique of all kinds of public choice architecture. In fact, my extended argument defends a particular, fairly extensive account of permissible public nudging – captured by what I call the *transformative nudge model* – which recognises and seeks to justify the agency-enhancing potential of choice architecture with respect to both self-regarding and other-regarding behaviour. Importantly, this political-theoretical argument is *empirically grounded*. It takes seriously the discoveries from the cognitive and behavioural sciences and examines how they threaten core liberal values and central components of human flourishing, before exploring whether public choice architecture holds out any promise of counteracting these negative effects. The first part of this subsection has offered an argument "in defence of theory" (Dworkin 1997: 375); this second part clarifies what type of theory I think it is that we should pursue in answering the question at hand.

Some believe that political theory should be situated within the realm of idealised abstraction, where its practitioners can conduct principled reflection into the fundamental principles of justice, which are taken to be logically independent both of issues of feasibility and of questions relating to human nature. This is not the method I adopt; I wish to engage in political theory with the express aim of finding principles to guide action in our current circumstances.[13] And, in my view, if political theory is to be action-guiding (which I take to be at least part of its function), theorists can no longer pursue their goals with blithe disregard for the empirical findings which raise essential questions about human agency (§2.1 presents this research). Since political theory is widely accepted as a practical discipline, it is possible to maintain both that empirical findings do not have a bearing on deciding upon fundamental political principles and that normative theorising should take account of relevant empirical findings when seeking to operationalise these fundamental principles for the guidance of action, given that the feasibility of implementing the principles alters in light of such facts. On this view, then, political theory informed by relevant empirical findings provides the most promising approach to producing theory – in this case, a theory of the political morality of public nudging – that is able to guide action towards our ideals here and now (Niker

[13] For more on this methodological debate between "ideal theory" and "nonideal theory", as well as a helpful discussion of the different kinds of nonideal theory, see Stemplowska and Swift (2012). For defences of the non-ideal approach, see Farrelly (2007) and Stears (2005).

2013: 22).[14] This openness to interdisciplinarity does, at certain points in the thesis at least, have important implications for my extended argument, which contributes to its distinctiveness relative to traditional autonomy-based liberal views. This point leads us into my second set of remarks about the methods I adopt.

§1.2.2: Living Well and the Demands of Political Morality

There are competing theories of the demands of political morality, and one's views on this set of foundational issues concerning political justification will form the theoretical premises from which one develops an account of permissible nudging. One might, for example, offer an account based on an acceptance of Rawlsian political liberal premises, which would determine the shape of the view in particular ways (e.g., see Moles 2015). But, as Moles points out, it is clearly the case that "not all liberals endorse this view" (2015: 645, fn. 5). This will be the case for any theory of the permissibility conditions for nudging, since normative theory presupposes some theory of value from which it derives the requirements that it imposes on the behaviour of individuals and states (Raz 1999: 11). The aim of this subsection, then, is to set out briefly the theoretical premises on which the extended argument of this thesis rests, in particular certain claims about what it means to live well and the role that a liberal state might justifiably play in supporting its citizens' in living well. To be sure, this will not amount to a defence of this view relative to its alternatives. Nonetheless, it will serve the useful function of situating my account within the broader theoretical terrain, as well as provide an opportunity to outline some of its attractions.[15]

The debate over the demands of liberal political morality is framed by two questions: one concerning theoretical method, and a second relating to political substance (Mulhall and Swift 1996: 251-252; Beckman 2001: 4-10; Quong 2010: 15-22). The first concerns how we should conceive of the relationship between ethics and politics in political theorising, which asks whether liberal political philosophy must be grounded in some particular ethical ideal. The key distinction here is between comprehensive and purely political conceptions.[16] The second relates to what we take to be the proper conduct of politics, and asks whether liberalism approves of a state promoting certain activities, ideals, or ways of life. Here, the key distinction is between perfectionist and anti-perfectionist conceptions.

[14] I outline my views on the relationship between political philosophy and public policy in further detail in §6.1.
[15] Note that there is a fuller discussion of these issues in Chapter 3.
[16] This terminology is introduced by Rawls (1993: 12-13).

With respect to theoretical method, I keep company with theorists like J.S. Mill, Ronald Dworkin, and Joseph Raz in assuming a *comprehensive* conception, which rejects the methodological axiom – central to Rawls's political liberalism – that political theorising requires disconnecting ethical ideals from politics.[17] In more positive terms, comprehensive conceptions maintain that (certain) ethical ideals are relevant to theorising about the proper structure and concerns of the sphere of politics. What unites them is a commitment to the idea that liberal political morality is grounded in a distinctively liberal conception of what constitutes living well. This is commonly based on the idea that, whatever else it might include, a liberal ethic assumes personal autonomy (or similar concepts, such as individual integrity) as a – if not, *the* – central component. In line with this, autonomy plays this central justificatory role in my account of permissible nudging.

Comprehensive conceptions can take either perfectionist or anti-perfectionist forms, depending on one's answer to whether it is permissible for states to actively promote living well. Perfectionism in politics is based on the idea, which is accepted as intuitively compelling even by its critics, that "the aim of the state (or at least one of its major aims) should be to improve the lives of citizens" (Quong 2010: 30). As a result, and in contrast to its rival position, it holds that "there is no fundamental principled inhibition on governments acting for any valid moral reason" when going about seeking to achieve this aim (Raz 1989: 1230). My account takes the form of a *perfectionist* political morality, because it accepts the following claim:

> *The liberal perfectionist thesis*: "It is at least sometimes permissible for a liberal state to promote or discourage particular activities, ideals, or ways of life on grounds relating to their inherent or intrinsic value." (Quong 2010: 27)

In particular, the account I develop argues that public nudging can, under certain conditions, be justified by reference to its ability to promote citizens' autonomy, rightly understood, on grounds that this has intrinsic value for their lives. In such cases, this political action would represent a form of *noncoercive* state perfectionism. This general idea is described by Steven Wall thus:

> "Noncoercive state perfectionist measures may be able to counteract or cancel various pressures and influences that would otherwise impede rational decision making by its citizens. Designed well, such measures might contribute to an ethical environment in

[17] For an account of permissible nudging based on Rawls' political liberalism, see Moles (2015).

> which people were best able to respond to the reasons for and against the options that they confront. If [so], then they may not invade, but rather protect and promote, autonomous decision making." (Wall 2012)

Those who favour anti-perfectionism might respond by arguing that there is a potentially insulting, even manipulative, character to this kind of state involvement in promoting well-being. According to comprehensive anti-perfectionists, a commitment to an autonomy-based liberal ethic necessitates anti-perfectionism: the idea, in short, is that the best way to promote this liberal conception of living well is for the state to remain substantively neutral, in aim, between all (reasonable) conceptions of the good.[18] This issue will be discussed in detail in subsequent chapters; but for now it is worth noticing that noncoercive state perfectionism might take one of two forms. It might be designed to protect and promote citizens' autonomous agency or to help citizens pursue or engage with certain worthwhile or valuable options (Wall 1998: 197-198). According to the account of the nature and value of autonomy I set out later in the thesis, permissible noncoercive (nudge) perfectionism can take only the first of these forms, in the prudential realm at least. As I will seek to show, my account is not subject to the criticisms often levelled at perfectionist political action by anti-perfectionists: the nudge interventions it deems permissible are not insulting, manipulative, or objectionably paternalistic. In general, they are motivated by an ambition – which, in light of the empirical evidence, is in some cases best characterised as a positive duty of governments – to support citizens' practical reasoning processes.

Some of these practical reasoning processes relate to citizens' self-regarding actions, while others relate to their other-regarding action. Accordingly, the normative analysis presented in this thesis is divided into two main strands, paralleling the two main sets of reasons why the political community might modify citizens' behaviour. The first is to improve the target's own life (let's label this *prudential nudging*), and the second is to improve the lives of others, by preventing harm or wrong to others (we can call this *moral nudging*). These two kinds of reasons activate different sets of normative concerns, and so it is necessary to assess these issues separately if we are to do justice to all of the relevant considerations. As will become more apparent over the course of the thesis, simply drawing this distinction and carrying out full examinations of the

[18] Since the liberal ethic offers a general philosophical conception of living well – i.e., attributing fundamental importance to the human capacity for autonomy – rather than more concrete statements about how people should exercise this autonomous choice, the comprehensive anti-perfectionist's state is justified by appeal to a conception of the good life in the first sense only, and is therefore able to remain neutral between conceptions of the good in the second sense (Mulhall and Swift 1996: 254).

normative dimensions of each type of nudging represents an important contribution to the literature in itself. The most novel and ambitious aspect of my analysis, though, is to offer a *complete* account, by which I mean an account that unifies the two strands of analysis – prudential and moral – in a coherent overall theory.

To understand the basis of this view, it is important to recognise that a moral nudge can in principle be done *for the sake of another* – this is what makes it moral in character – while still being partially motivated and/or justified by prudential or ethical concerns relating to the target of the nudge.[19] On my autonomy-based view of living well (outlined in Chapter 3), it matters not only that we discharge our duties and responsibilities to others, but that we do so autonomously. That is, there is ethical value in fulfilling a duty as a response to recognising the reason(s) for doing so, which is not the case when we fulfil a duty because, say, we were hypnotised or forced in some other way into doing so. Although these considerations are often outweighed, there is something distinctive about the case of public choice architecture and its unique promise in the moral realm – namely, its ability to help governments deal with problems associated with nonenforceability – that results in these ethical considerations taking on a special significance.

Broadly speaking, then, what drives and unites this normative investigation, in all its variety, is a commitment to the value of personal autonomy and its role in living well. Within a web of complexities, this simplicity counts as a virtue of the theory I defend. This is because it is a desirable feature of a normative theory, to quote Shelly Kagan, that "it yields a body of judgments out of a relatively sparse amount of theory, deriving the numerous complex variations of the phenomena from a small number of basic principles" (1989: 11; cited in Parr 2015: 24-25). To be sure, this can be a virtue of my theory only if the analysis it generates is as compelling as any alternative; however, he remainder of the thesis seeks to show this to be the case.

§1.3: Preview of the Thesis

We can now specify the thesis's central question in more specific terms in light of this premise: how, if at all, can governments permissibly apply the psychological evidence to

[19] In using "ethical" in this way, I follow Dworkin. There are different ways of understanding this term. Some use it in a broad sense, which effectively erases the distinction between "ethical" and "moral" (Dworkin 2011: 191). I would suggest that Sunstein understands it in this way when entitling his book *The Ethics of Influence*. In its narrower sense, however, "ethics" concerns how we ought to live our lives. This is distinct from "morality" since this prescribes how we ought to treat others. As I suggest here, however, it is possible – and indeed, I think it is true to some extent – that "our desire to lead good lives for ourselves provides a justifying reason for our concern with what we owe to others" (Dworkin 2011: 191).

the design of public policy and social institutions as a means of improving citizens' lives?

I begin to lay the foundations for answering this question in Chapter 2. Its first set of tasks is descriptive. First, it introduces the empirical findings from the cognitive and behavioural sciences that have "generated a radical change in the broader framework within which human nature is investigated" (Lavazza and De Caro 2010: 23), and explains how and why these findings have provided the basis for a new behavioural insight-led approach to government. Second, it outlines the three-part moral argument offered by Thaler and Sunstein (2009) that aims to justify this shift in the logic of policymaking. The chapter then engages critically with this argument and, in so doing, clarifies the character of nudging, before introducing the range of normative concerns that have been raised regarding the permissibility of this means of influence. Mapping out this critical terrain not only offers a review of the current literature, but allows me to highlight the gap that this thesis's more nuanced argument intends to fill.

What this discussion highlights is the need to critically examine the framework within which the current debate over the ethics of nudging is taking place. I do this in Chapter 3, where I argue that we should take an *autonomy-based*, rather than a preference satisfaction, approach to assessing whether nudges can improve the lives of citizens. The chapter begins by reconstructing the standard used by nudge advocates for determining what it is that makes a person's life go well – one that rests on the satisfaction of a person's rational preferences – and explains why the adoption of this standard is taken to be attractive (namely, that it appears to justify only means, and not ends, paternalism). It then outlines a competing approach, grounded in a partial account of well-being that understands personal autonomy to be a central component of what it means to live well. The second part of the chapter explores the theoretical implications of taking this approach, partly to diffuse some of the controversiality thought, by some, to be attached to the adoption of this standard. It does this in two ways: first, it clarifies the moderate character of my liberal perfectionist framework and questions whether, in fact, anti-perfectionists would necessarily reject the content of the account it generates; second, it highlights serious problems with the consistency of the informed preference satisfaction approach, which purports to set itself up in opposition to all forms of perfectionism. In this way, Chapters 2 and 3 serve to lay the conceptual and theoretical foundations for the normative inquiry that follows.

As already noted, there are two main strands to this normative analysis of public nudging, one relating to behaviour change policies designed for the sake of the target and the other relating to those designed for the sake of others. These analyses are performed in Chapters 4 and 5, respectively, which comprise the main argumentative body of the thesis. Chapter 4 develops and defends a selective account of permissible prudential nudging, which it calls *autonomy-supporting nudge paternalism*. It carries out a full examination of the way in which nudges operate on their target, drawing a series of important distinctions which enable us better to comprehend and to capture the ethical and political complexities involved in the use of this policy tool. As a result, the chapter engages in a more nuanced normative analysis than is often usual within the current debate, showing that certain types of nudges are manipulative and thus presumptively impermissible, while others are able to act as a form of autonomy support in virtue of their ability non-manipulatively to mitigate the internal threat to autonomy that comes from various cognitive biases. The chapter sets out two categories of choice architecture – *nudge-autonomy* and *auto-nudge* – that are able to play this autonomy-supporting role.

Chapter 5 assesses whether there are conditions under which it is permissible to use nudges as a means of moral improvement – something that has not received any attention in the debate thus far. After characterising moral improvement and considering the different ways we might justify governmental efforts to bring about this end, the chapter sets out a two-part account of the conditions under which the use of *moral nudges*, as I call them, is permissible. The first examines the conditions under which it would be desirable to use moral nudges – of the kind deemed impermissible in the prudential realm – to secure conformity with moral reasons in the case of enforceable duties. It sets out a restrictive, conditional view on when nudges might be preferable to (additional) coercive measures. The main promise of choice architecture, in my view, is to be found in helping governments to deal with problems associated with the nonenforceability of certain moral duties. To this end the second, and main, part of this investigation develops a detailed account of when moral nudges can be used in this capacity. It sets out three kinds of interventions – *motivational scaffolding*, *inducing moral deliberation*, and *public-ecological persuasion* – that are deemed permissible in part because they support agents in developing and exercising autonomous (and, in this case, morally responsible) agency, which has ethical value for them as well as the moral value that comes from preventing harm to others.

Chapter 6 aims to translate the lessons from this comprehensive normative analysis into an account of permissible public nudging in practice. This engagement with the practical concerns and issues relating to the making of behavioural policy marks a key contribution of the thesis. The chapter begins by articulating the essential form of the thesis's extended philosophical argument, which it expresses in two ethical principles that form of the basis of the *transformative nudge model*, namely, the *nudge-to-deliberate principle* and the *nudge-to-motivate principle*. Within the context of the UK and its Behavioural Insight Team in particular, it then explores the difference that the two ethical principles might make to existing governmental approaches to designing choice architecture here and now. It sets out an initial ethically-sensitive policy framework to guide the design of interventions, and considers some potential issues concerning its implementation. The chapter finishes with a discussion of some of the limitations of the thesis, picking out some of the main ways in which its analysis could be extended in further research. This is followed by a brief conclusion (in Chapter 7), which, rather than restating the thesis's arguments and main conclusions, offers some concluding remarks about my hopes for this debate going forward.

The thesis aims to offer a distinctive and timely contribution to the literature, as well as to the public debate over the ongoing development and use of this policy tool. Its main contribution to the literature is to present a complete account of the ethics of psychologically-informed public policymaking, which is able both to unmask and, more importantly, to solve in a coherent and theoretically attractive fashion a set of deep problems in the current debate. This account is distinctive in two significant ways. First, it differs from traditional autonomy-based accounts in the political theory literature because it takes seriously the empirical evidence from the cognitive and behavioural sciences and, in particular, the threat to autonomy associated with our automatic and context-dependent mental processes. It claims that those interested in autonomy should be interested in and attentive to this evidence, and defends an autonomy-based account of the permissibility of nudging that is appropriately updated in light of these findings. Second, as already noted, my account is distinctive in virtue of its completeness. Much of the critical engagement with nudging has been piecemeal and has often been constrained by the (sometimes problematic) contours set by influential early contributions. This thesis resists these limitations, which allows it to open up the debate in new and underexplored ways. In particular, the thesis presents the first full investigation into choice architecture's potential moralising effects and the role that they

might play within a theory of the political morality of public nudging. The account I offer is not only distinctive, but politically timely. It represents the first attempt, to the best of my knowledge, to integrate serious philosophical assessment into existing practices in behavioural policymaking. At a time when the vast expansion of behavioural policy is raising questions about its ethical dimensions, even among its practitioners, an account that is able to translate its arguments into simple ethical principles is precisely the kind of contribution that is called for, and my hope is that it will open up a productive, two-way discussion about developing an ethically-sensitive framework for designing and implementing public nudges.

2. New Models of Agency and Influence

A large and ever-growing body of multidisciplinary research has revealed that classic models of human agency are in serious need of revision. Understanding this shift in the empirical foundations of human behaviour and decision-making is important for our normative inquiry because it has provided the basis for a moral argument in favour of changing the way we think about what counts as permissible governmental influence, as well as the main practical insights into what form these new behavioural policy levers might take. The first aim of this chapter, then, is to outline both the relevant empirical evidence (§2.1) and the argument for nudging (§2.2.1). Following this, its second aim is to critically engage with this argument in order to clarify the character of nudging as I will understand it in the remainder of the thesis (§2.2.2), before mapping the existing critical terrain (§2.3).

§2.1: Human Agency and the "Socio-Ecological" Turn

Despite being a complex and dynamic field of investigation within the cognitive and behavioural sciences, there is an "emerging *consensus* about the fundamental point relating to both the scope and limits of human rationality and the cognitive architecture that supports it" (Samuels et al. 2012: 192; emphasis in original). As a result of this consensus, a new model of human agency, based on dual-process theories of cognition, has risen to prominence – one that is able to make sense of the 'anomalies' that occur as a consequence of our automatic and context-sensitive mental processes, as well as to explain how these processes are related to the more familiar deliberative processes that have hitherto taken centre-stage in our theories of rational agency.

This new model has been met with different types of responses in the literature: some see it as having highlighted that we are subject to "truly dismaying" limitations as thinkers (Levy 2012: 589), while others interpret it as revealing more about how the human mind achieves its "unique brilliance" (Haidt 2006: 17). However we might think of it, this alternative model has transformed our understanding of the "socio-ecological" factors acting on us (Hurley 2011), and has revealed new spaces and modes of influence that previously had gone either under-appreciated or unrecognised. This section presents an overview of the findings that have led to this new model of agency (§2.1.1), and then discusses the transition to thinking about these findings as "behavioural insights" with relevance to policymaking (§2.1.2).

§2.1.1: Dual-Process Theories of Cognition

Robust empirical evidence has provided the basis for an "emerging analytics of decision making" that diverges from classical accounts of individual rational agency in significant ways (Whitehead et al. 2011: 2819-2820). Two key assumptions tend to unite classical conceptions: first, that rational agency is marked by a person's perceptions being *decoupled* from her behaviour, such that the former play "flexible, instrumental roles in relation to action, mediated by the individual's desires and ends"; second, that rational agency is *domain-general*, by which is meant that it is not tied to specific domains or environments (Hurley 2011: 194-195). These assumptions have provided the foundation for the economic theories and policy recommendations by which we have made sense of and governed our social world, almost as if they were basic laws of human nature (Ariely 2009). However, evidence relating to the phenomena of *automaticity* and *context-dependency* has called each of these assumptions into question, respectively, and has been at the heart of forging an updated conception of human cognition and agency.

Cognitive decoupling is taken to be an essential part of the kind of conscious process that is necessary for a person to judge which action, from the available set of options, is (or is believed to be) in his or her best interest or is (or is believed to be) morally appropriate to the situation at hand. If perception and behaviour are decoupled, the supposed result is that behaviour is necessarily mediated by conscious processes. An action would be the result of a two-stage process: first, a person takes in perceptions and, through internal reasoning processes, arrives at beliefs about the world including beliefs about the probable results of alternative actions; second, these alternative actions are valued instrumentally by reference to the person's ends, thereby producing a pattern of choices that is consistent and instrumentally effective relative to his or her beliefs and ends (Hurley 2011: 195). Research into automaticity has challenged the validity of this two-stage process as a general assumption about human decision-making. While its early beginnings can be found in Williams James' ideomotor theory (1890), the phenomenon of automaticity has more recently been a central feature of the work of social psychologist John Bargh and his colleagues, marking out a core research agenda for scientists of human behaviour. These researchers have found that decoupling is not representative of how behaviour comes about within the social contexts in which we live much of our lives. Rather, our behaviour can be triggered automatically by the mere presence of relevant situational features in a way that is not mediated by the kinds of

conscious perceptual and judgemental processes described in the two-stage process. This has been labelled the "perception-behaviour link" (Bargh et al. 1996: 231-233).

Consider *priming effects* as an illustrative example. This refers to the implicit memory effects that occur without any awareness, in which exposure to a particular stimulus nonconsciously "primes" a response to a later stimulus. In one study, participants were asked to unscramble sets of five words and make sentences using four of them (e.g., you might make the sentence 'he finds it instantly' from the word-set 'finds he it yellow instantly'). Half of the participants were given sets that contained words associated with the elderly (e.g., 'wrinkle', 'forgetful', 'grey', 'bingo', 'bald', etc.), while the sets for the other half did not contain any of these words. Following completion of the task, participants were asked to walk down the corridor to complete a second experiment. Unbeknownst to them, the experimenters were recording the time it took them to walk to the other room. Those participants who had been exposed to words relating to the elderly walked significantly more slowly than the control group (Bargh et al. 1996). In a similar experiment, whether a participant unscrambled words that had associations with rudeness or politeness was shown to have an effect on how she related to the experimenter to whom she had been asked to report, but who was engaged in a conversation with someone and was not making any eye contact with her. Even more interesting was the finding that these effects are not dependent upon people consciously reading the words; the same effects can be shown to occur when the words are presented to participants subliminally, i.e., by flashing them on a screen for just a few hundredths of a second, thereby ruling out the possibility that they will be registered consciously (Haidt 2006: 14). This indicates that some part of the mind nevertheless does register the words, and that this sets in motion the behaviours measured in the studies, such as walking slowly or interrupting the experimenter from her conversation.

How should we understand these results, and the many others like them? The pervasive response has been for many psychologists and neuroscientists to converge in support of *dual-process theories of cognition* (Chaiken and Trope 1999; Stanovich and West 2000; Evans 2003; Kahneman 2003: 1450-1452). Such theories rest on the claim that human reasoning is comprised of two main types of processes, and they often use the generic terms "System 1" and "System 2" to label their distinct sets of properties (Stanovich 1999). The attributes of System 1 have been characterised variously as: automatic, fast, unintentional, nonconscious, associative, contextualised, intuitive,

uncontrolled, implicit, subject to biases, evolutionarily ancient, resistant to interference from concurrent processes, and tending to link perception fairly directly to action. By contrast, System 2 has been characterised as: reflective, slow, deliberate, intentional, conscious, controlled, flexible, explicit, uniquely human, effortful, costly in terms of cognitive and attentional resources, analytic and critical, domain-general, subject to interference from other cognitive processes and to disruption by time pressure or cognitive load, tending to decouple perception from behaviour (Evans and Stanovich 2013: 225, see Table 1; Hurley 2011: 203). Particular theories may foreground different elements of this broad contrast, though these family disputes are not relevant for our purposes; what is important is that there is a clear consensus on the dual-process account as representative of the nature of the cognitive architecture supporting human reasoning.[20]

A number of metaphors have been used to help us think about the divided nature of our mind and its reasoning processes.[21] Daniel Kahneman, for instance, opens *Thinking, Fast and Slow* by describing the book as a "psychodrama" with two characters – the automatic System 1 and the effortful System 2 – whom he invites us to think about as agents with "their individual abilities, limitations, and functions" (2011: 21).[22] We identify ourselves with the conscious and deliberative agent, represented by System 2; but Kahneman describes System 1 as the story's hero, "effortlessly originating impressions and feelings that are the main sources of the explicit beliefs and deliberate choices of System 2" (ibid.). Even so, much of the book is devoted to understanding the "uneasy interaction" between these two systems and, specifically, the systematic errors of intuitive judgement and choice (i.e., the cognitive biases) that he attributes to the workings of System 1 (ibid.: 415-416). More creatively, Jonathan Haidt describes human reasoning in terms of an elephant and its rider. The rider represents the conscious, deliberative processes of System 2; while the elephant represents everything else: "the gut feelings, visceral reactions, emotions, and intuitions that comprise much of the automatic system" (Haidt 2006: 17). Each has its own kind of intelligence; and the unique brilliance of human beings is made possible by these systems, and their

[20] For a good overview of the current state of the debate on dual-process theories, see Evans and Stanovich (2013).

[21] There is a good psychological reason for this, as Haidt explains: "Human thinking depends on metaphor. We understand new or complex things in relation to things we already know" (2006: 2). For more on what they call "the neural basis of metaphorical thought", see Lakoff and Johnson (2003: 257-261).

[22] He also makes explicit reference, for the sake of clarification, to the metaphorical nature of these agents: "I describe mental life by the metaphor of two agents, called System 1 and System 2, which respectively produce fast and slow thinking" (Kahneman 2011: 13).

respective intelligences, working together in synergy. The problem is that they do not always work together in this way: "Like the rider on the back of an elephant, the conscious, reasoning part of the mind has only limited control of what the elephant does" (ibid.: xi).

It is at this point that we can begin to understand why the second assumption of classical conceptions of rational agency, domain-generality, is also found to be wanting. It is the interplay of nonconscious cognitive heuristics with situational features that ends up being "the secret author of many of the choices and judgments [we] make", despite the fact that, from the perspective of classical conceptions of rationality, these often very minor contextual details should be irrelevant to the outcome (Kahneman 2011: 13; also see Simon 1990: 7). Cognitive and evolutionary psychologists have discovered an extensive set of heuristics that operate as mental shortcuts (or information-processing rules), while social psychologists and those working in the field of cognitive engineering have amassed a vast amount of evidence on how the structure of environments interact with and modulate this "fast thinking" (Norman 2013).[23] Some choice environments work with the grain of these automatic processes, as it were, so as to produce good (or good-enough) behavioural outcomes; but others generate *cognitive biases*, i.e., processes that are considered as nonrational on outcome- and process-based views of classical rational agency.[24]

The study of biases has become an extensive area of research and, increasingly, biases are both gathering public attention and having an impact on disciplines outside of the cognitive sciences. Within medical practice, there is an ongoing debate about the role of *framing effects* on the decisions of both patients and doctors (Marteau 1989; Gigerenzer 2014: 159-224). Patient's preferences and doctors' judgements are sensitive to the way in which information is framed: "Even experienced medical doctors are more likely to choose a treatment when its effects are described as 200 lives saved out of 600, as opposed to 400 deaths out of 600" (Moles 2015: 646; Kahneman 2003: 1458). Within policy, much attention has been given to the impact of *status quo bias* or *default*

[23] Tversky and Kahneman are generally considered the fathers of the "heuristics and biases" tradition. In a famous paper published in 1974, they argued that three heuristics underlie a wide range of intuitive judgements: the *representative heuristic* (which states that probabilities are evaluated by the degree to which X resembles Y); the *availability heuristic* (which captures the fact that we estimate likelihood by the ease of mental association or retrieval); and the *anchoring heuristic* (which captures the fact that different starting points yield different estimates that are biased towards the initial values) (Tversky and Kahneman 1974; Blumenthal-Barby 2016: 5). Following decades of research since then, the current set of biases and heuristics can no longer fit neatly into these three initial categories. Indeed, in a recent review of 214 empirical studies of biases and heuristics, 19 different types were identified (Blumenthal-Barby and Krieger 2015; for a list and short description of these, see Blumenthal-Barby 2016: 6, Table 1).

[24] For a discussion of these two forms of the classical conception, see Hurley (2011: 193-194).

bias. People have a strong tendency, whenever a default is offered, to stick with this. This is starkly illustrated, for example, by the difference between two similar European countries with respect to organ donation registration: in Germany, which has an opt-in system, only 12% of citizens have given consent; whereas in Austria, which uses a presumed consent system, 99% of people are on the register. Within the legal and criminal justice systems, as well as in philosophy, there is an important discussion about *implicit stereotype biases*. People often manifest implicit attitudes towards members of socially stigmatised groups on account of unconscious stereotype associations, and do so even when they hold explicit anti-discriminatory or egalitarian beliefs about that group (Brownstein 2015). Implicit biases therefore have serious implications for society, especially as regards discrimination, prejudice, and structural injustice.[25]

Dual-process theories offer a means of explaining the systematic behavioural and cognitive anomalies – from the perspective of classical conceptions of individual rational agency – that are part and parcel of our everyday experience. Of course, dual-process theories are also able to explain why we are able effectively to deliberate and make decisions in line with our goals, preferences, and values, especially when we are not in conditions that put pressure on our System 2 processes. Nevertheless, the evidence has revealed that human agency is more of a socio-ecological phenomenon than we have traditionally thought; under certain circumstances, our behaviour tends towards being "embedded in and dependent upon [features of our] social environment" (Hurley 2011: 192). It is this predictability, albeit probabilistic in nature, that has made possible (and necessary, according to some) the move towards behavioural theories of economics, as well as providing the foundations for an important shift towards the use of knowledge of these biases at the level of public policy and governance – the particular development that interests us here. Within this domain, these biases have become more commonly known as "behavioural insights".

§2.1.2: Biases as "Behavioural Insights"

Governments have various tools at their disposal for influencing and intervening in citizens' behaviour in accordance with their legitimate aims (see House of Lords 2011: 10, Table 1). Knowledge of predictable cognitive biases has revealed the possibility of adding another policy lever to this list. The behaviour change agenda is based on the

[25] For a discussion of the impact of cognitive biases on criminal justice, see Benforado (2015); and for a discussion of the mechanisms available within British law in particular for working against the effects of implicit bias, see Hosking and Russell (2016). Also, see Brownstein and Saul (2016) for a collection of philosophical contributions concerning the relationship between implicit bias and structural injustice.

idea that these lessons from psychology and the cognitive sciences can be applied by governments to the design of public policy in ways that can effectively influence and modify citizens' behaviour without coercing, incentivising, or educating them, as such. Such efforts have been described using a range of terms: "behavioural policy" (Halpern 2015), "institutional prosthetics" (Trout 2005: 394), "ecological engineering" (Levy 2012) and, most commonly, "nudging" or "choice architecture" (Thaler and Sunstein 2009; Sunstein 2014a).[26]

Two key claims unite these proposals. First, given that human behaviour has been shown to be context-sensitive under certain conditions and in certain domains, it is possible to design public environments that are more or less conducive to generating particular behaviours. This rests on the possibility of making use of behavioural insights primarily as a form of "nonargumentative influence", namely:

> "…influence that operates either by bypassing a person's awareness or by relying on facts about the subject's psychology.... It is in contrast to influence that operates by […] offering him reasons and arguments (i.e., rational persuasion), and to influence that operates by force or severe threats of harm (i.e., coercion)." (Blumenthal-Barby 2014: 123)

Second, now that we are aware of this possibility, we ought to design these public environments in ways that work with the grain of human psychology, for the purpose of making citizens' lives go better. The first claim is an empirical one, while the second is normative and introduces the issue of paternalism to the debate. Indeed, from both a policy and a political-theoretical perspective, the main impact of the discoveries in cognitive sciences has been to challenge the traditional philosophical animus against paternalism, and to argue in favour of a more diversified picture. Those who claim that some forms of paternalistic intervention are permissible, and sometimes even morally

[26] Some of these terms, especially those making reference to behavioural engineering, may strike one as somewhat behaviorist in character (e.g. Skinner 1971). There is an important fundamental sense is which these proposals differ from the behaviourist project, however. As Graham (2015) explains, the doctrine of behaviourism is committed to two claims: (1) that psychology is the science of behaviour, not the science of the mind; and (2) that behaviour "can be described and explained without making ultimate reference to mental events or to internal psychological processes", that is, that the sources of behaviour are to be found in the environment, not in the mind. By contrast, the political proposals based on dual-process theories of cognition reject these claims and instead focus on taking advantage of what we know about the relationship between mental processes and environmental factors (for a helpful discussion of Skinnerian behavioural control, see Dworkin 1988: 150-160). Even though this critique is misplaced in formal terms, it does capture an important worry about these kinds of behaviour change techniques – namely, the worry about manipulation and the undermining of autonomy more generally – that are discussed in detail in subsequent chapters.

required, are no longer rarities.[27] These proposals have not been confined only to "soft", understood as noncoercive, forms of paternalism.[28] Some have used the evidence as a basis for offering a philosophical defence of coercive paternalism, i.e., "paternalism of the sort that forces people to act, or refrain from acting, according to their best interests" (Conly 2013: 3). These kinds of arguments, especially those advocating for soft versions of paternalism, have already had a significant impact on policymaking (as we saw in §1.1) – and none more so than Thaler and Sunstein's case in favour of what they call "libertarian paternalism", which promises not to offend against citizens' freedom of choice.

§2.2: From Cognitive Architecture to Choice Architecture

Libertarian paternalism is arguably the most theoretically innovative of the proposals in this new wave of purportedly permissible paternalistic influence. But before proceeding in discussing it further, it is worth noting a terminological point. In his Foreword to David Halpern's book *Inside the Nudge Unit*, Richard Thaler points out a difference in how this proposal is labelled in US and UK contexts. He writes: "We called our philosophy libertarian (or liberal in the UK) paternalism" (Thaler 2015: x). In what follows, then, I will understand their argument as one seeking to justify a distinctively liberal form of paternalism, and will drop the reference to libertarianism. I will instead refer to their proposal as *nudge paternalism*. This section has two tasks: first, it presents Thaler and Sunstein's argument for nudge paternalism (§2.2.1); second, it offers some initial critical engagement with this argument, in order to highlight some points that are important to understanding the character of nudging (§2.2.2).

§2.2.1: The Argument for Nudge Paternalism

Thaler and Sunstein's argument is based on three claims: (1) that the empirical "best judge" assumption is false; (2) that paternalism, in some form, is inevitable; and (3) that paternalism need not be coercive. The first follows directly from the empirical research

[27] For example, in addition to Thaler and Sunstein (2009) on "libertarian paternalism", see Camerer et al. (2003) on "asymmetric paternalism" and Ahlstrom-Vij (2013) on "epistemic paternalism".

[28] There are two ways in which the distinction between "soft" and "hard" paternalism is used in the literature. The first parallels the distinction between noncoercive and coercive paternalism. Thaler and Sunstein claim that their proposal is a form of soft paternalism in this sense (although this has sometimes been contested, as we will see below). The second, arguably more common, way is outlined by Feinberg (1989) and focuses on voluntariness. On this view, soft paternalism maintains that a government has the right to intervene "when but only when [a person's] conduct is substantially nonvoluntary, or when temporary intervention is necessary to establish whether it is voluntary or not" (Feinberg 1989: 12). It is less clear that nudge paternalism is soft in this second sense. We will discuss issues pertaining to this in the next chapter.

outlined above.[29] It is a denial of J.S. Mill's so-called epistemic argument, which holds that individuals are in the best position to identify both their own ends and the best means of obtaining them, because they have "means of knowledge immeasurably surpassing those that can be possessed by anyone else" (Mill 2005: 93; see Sunstein 2014: 6-7). Thus, if the government's goal is to ensure that people's lives go well, allowing people to pick out their own paths offers the most effective route.[30] Sunstein contends that this argument, based on the non-normative *best judge assumption*, "provides the strongest support that the Harm Principle can find" (2014: 7); but, he also argues that it is no longer available to anti-paternalists on account of the fact that people systematically make nonrational choices in certain contexts that can harm their interests in more or less severe ways.

The second claim has come to be known as the *argument from unavoidability*, which many view as Thaler and Sunstein's "most important argument for nudging" (Grill 2014: 142). It purports to show that the idea that there is an alternative to at least some of kind of paternalism, which is a necessary premise or assumption of anti-paternalism, is a misconception. The central thought is that the research highlighting the context-dependency or endogeneity of people's preferences demonstrates that "*Choice architecture is inevitable*" (Sunstein 2014: 118 and 2016a: 76; emphasis in originals). Thaler and Sunstein open *Nudge* with the motivating illustrative example of Carolyn, the director of a school (or university) cafeteria (2009: 1-2).[31] Carolyn notices that her arrangement of the lunch items in the canteen has a significant determining influence on what students choose; in particular, the majority select the option that is presented first. Once noticed, this power to influence can be exercised in different ways. She could use it to maximise cafeteria profits, or to direct students towards healthier options, or she could use a randomising technique in an attempt to neutralise her power. What cannot be changed, however, is the fact that the items must be presented in *some* order, which means that

[29] These authors are clear that the "basic source of information" used to formulate and justify nudge interventions is the recent evidence about human decision-making outlined in the previous section (2009: 8). They interpret it as highlighting that decisions that follow from cognitive biases are "flawed" and "systematically wrong" judgements, which result from "human frailty" (ibid.: 19, 5, 79).

[30] According to Mill, this is because: "The interference of society to overrule his judgement and purposes in what only regards himself must be grounded on general presumptions which may be altogether wrong and, even if right, are as likely as not to be misapplied to individual cases, by persons no better acquainted with the circumstances of such cases than those are who look at them merely from without" (2005: 93).

[31] Their use of an example that involves children is problematic, because we generally consider there to be different principles that apply in the cases of children and adults on account of the former's lack of autonomy (for more on this, see §6.4.1). This rhetorical strategy of Thaler and Sunstein's has been called out in the critical literature. Here, then, I suggest modifying the example so that the students are students at a university, rather than at a school.

Carolyn assumes the role of "choice architect".[32] If we extend this example more generally, it demonstrates that it is impossible for a government to avoid influencing its citizens: it is, simply by virtue of its legitimate role and status, inevitably involved in the business of structuring the "landscape of choice" in which its citizens are embedded and in which they navigate their lives (Ben-Porath 2010). According to Thaler and Sunstein, this highlights that there is "no way of avoiding nudging in some direction" (2009: 5). Like Carolyn, government actors are choice architects; once we realise this, we should accept that these actors have good reasons to design policies and institutions in ways that have beneficial effects for citizens. This argument purports to show that paternalism, in some form and in some contexts at least, is *inevitable*.

The third claim is opens the door to the possibility of a distinctively *liberal* type of paternalism, which is intended to be acceptable to liberals of all persuasions. This involves recognising that coercive paternalism, that is, interventions that regulate the landscape of choice in ways that simply remove certain options or significantly increase their costs, is not the only route available to us. Choice architecture, it is argued, offers a *noncoercive* form of paternalism. We might purposefully design choice environments in ways that utilise what we know about cognitive biases and heuristics to direct people towards "welfare-promoting" behaviour (the paternalistic aspect); but, unlike its traditional counterpart, this form of governmental intervention is noncoercive in so far as it leaves all of the pre-existing options open and does not compromise people's freedom to choose between these options (the liberal aspect). In the cafeteria case, for example, lunch items might be arranged with a paternalistic motivation, without forcing a particular diet on anyone (Thaler and Sunstein 2009: 11). This type of intervention is "liberty-preserving", so the argument goes, because it does not "block, fence off, or significantly burden" any of the options (ibid.: 5). Since all of the options are still available, it would be easy for someone to resist, or to "opt out" of, the paternalistic arrangement. Consequently, nudge paternalism represents "a relatively weak, soft, and nonintrusive type of paternalism" that should be acceptable to liberals (ibid.).

Taken together, nudge advocates maintain that, in contexts where people are prone to error and where paternalistic intervention would improve their welfare, these three claims provide us with "a strong and emphatically *moral* argument on behalf of paternalism" (Sunstein 2014: 5; emphasis in original). This paternalism takes the form of "initiatives that maintain freedom of choice while also steering people's decisions in the

[32] On Thaler and Sunstein's view, she assumes this position whether or not she is aware of this power to influence. As we will see below, this expansive view should not be part of the characterisation of nudging.

right direction (as judged by themselves)" (ibid.: 17). The central idea is that, through situational design, nudges can promise an effective way for a government to help its citizens access the welfare gains of paternalism without incurring the losses to freedom which this usually entails. As already seen, this argument has had a significant impact on policymaking and has played an instrumental role in destabilising the philosophical animus against paternalism. It is, however, philosophically imprecise in various ways. There are problematic aspects about each of its three parts. I outline these below, with the purpose of clarifying the characterisation of nudging I will be using in the thesis. In some cases, this will also serve to flag issues that will be taken up in greater detail in later chapters.

§2.2.2: Clarifying the Character of Nudging

There are three main points of contention. These relate to: (i) the inevitability of choice architecture; (ii) the confounding of nudging and paternalism; and (iii) the purported "liberty-preserving" nature of choice architecture.

The first challenges how we should understand the claim of inevitability, which is generally taken to be the "core" of this ethical defence (Johnson 2016). There is a lack of clarity in Thaler and Sunstein's writings about whether intentionality is required for a policy design feature to count as an instance of choice architecture. At times, they refer to "unintentional nudges", suggesting that nudging can come in intended or unintended forms (Thaler and Sunstein 2009: 11). They downplay this distinction, perhaps because doing so provides their argument from unavoidability with added force. But choice architecture is not inevitable. To see this, we need to recognise an important distinction between *choice architecture* and *choice environments*. Governments might inevitably be involved in structuring the "landscape of choice"; but there is a difference between a choice environment being created the way it is (e.g., all public policies affect the choice environment in some way), and its being intentionally created the way it is *as a choice environment*. Only the latter can appropriately be called choice architecture, that is, the behaviourally-informed "curation" of public choice environments, by a choice architect, for the sake of bringing about a particular kind of behaviour change (Jennings et al. 2016). As a result, governments could inevitably be involved in structuring choice environments, without being unavoidably involved in the activity of choice architecture. On my characterisation of nudging, then, intentionality is a necessary component of choice architecture. This intentionality is at the heart of many of the objections to nudging. As one commentator points out, "what critics are concerned with is typically

not that people are influenced by their choice contexts, but that their choice contexts are designed, by others, with the intention of producing some particular behaviour" (Grill 2014: 143).

However, this type of intentionality need not be morally problematic in itself: that depends on the ends toward which this means of influence is directed. A second kind of problem, then, arises because Thaler and Sunstein's argument does not claim only that choice architecture is unavoidable, but that, as a result, some form of *paternalism* is inevitable. This confounds nudging, a particular means of influence, with nudge paternalism, a particular kind of moral justification for the governmental use of this influence. This is partly due to Thaler and Sunstein's definition of paternalism, which is "very weak in the sense that it allows many more acts to count as paternalistic than would be under almost all traditional definitions of paternalism" (Dworkin 2017). In essence, nudges are paternalistic, on their view, because they seek to promote the good of the person who is the target of the intervention.[33] Part of the issue of whether nudges are paternalistic in a more robust sense rests on a further issue, discussed below, concerning nudging's liberal credentials. The point I want to make here is that, even on Thaler and Sunstein's own definition, there are *nonpaternalistic nudges*, which shows that there is no necessary link between choice architecture and paternalism. Indeed it has always been the case that some of the prominent examples of nudge interventions do not fit neatly, or at all, into this overarching normative policy programme within which nudge advocates have embedded them. For example, charitable giving was one of earliest applications in the work of the UK's Behavioural Insights Team (Cabinet Office 2013; BIT 2015: 37-39; BIT 2016: 50-54). Another example is nudges that change the default rule for organ donation registration to increase the supply of available organs. In both cases, it is not the target of the nudge, but rather third-parties, who ultimately benefit.[34] This shows that there is nothing inherently paternalistic about nudging.[35]

---

[33] This might, more correctly, be defined as benevolence. Gerald Dworkin (2017), one of the foremost scholars of paternalism, defines it as "the interference of a state or an individual with another person, against their will, and defended or motivated by a claim that the person interfered with will be better off or protected from harm".

[34] There is, nonetheless, a difference between these two types of nonpaternalistic nudges. This relates to whether or not the individual who is the target of the intervention can reasonably be described as sharing in the collective benefit generated by the policy (see Raihani 2013: 2; see Table 1). In the case of nudges that seek to increase charitable donations by enrolling employees into an opt-out scheme, the employees cannot be said to share in the collective benefit, and there is a cost involved for them. In contrast, efforts that aim at increasing population health, such as increased participation in organ donor schemes, come with collective payoffs in which the target of the nonpaternalistic nudge might share (e.g., if she needed an organ transplant in some years' time, there would be more of chance that she will have access to this life-saving opportunity).

For this reason, the discussion of the permissibility of nudging is best conducted outside the narrow focus on libertarian paternalism that has characterised much of the debate so far (Schubert 2015: 3), which is something that Sunstein now admits (2016a: 54). Freeing the broader notion of *nudge-as-policy-lever* from the narrower conception of *nudge-as-libertarian-paternalism* is an important starting point for our investigation, because, like other policy levers, it may be permissible to employ choice architecture in the pursuit of a number of legitimate governmental goals. This is not yet to say anything about the conditions under which some sort of nudge paternalism should count as a legitimate aim – we will move on to this question, and other related ones, in subsequent chapters. For now, this point simply clarifies that nudging, as it is understood in this thesis, is a policy lever that need not necessarily be attached to any particular claims about the ends to which it can be used to serve.

The third issue relates to the claim that nudging is *liberty-preserving*. The attractive aspect of choice architecture for many is the idea that it preserves freedom of choice on account of the fact that it is designed to include an acceptably low opt-out cost. But it is not clear that all nudges meet this standard, even if some do. Assuming for now that freedom of choice is what matters from a liberal perspective (an assumption that is challenged below), there is still an important limitation relating to this so-called opt-out clause. To see this, we might distinguish between two kinds of freedom to choose: these have been described variously as "formal", "basic", or "nominal", on the one hand, and "effective", "substantive" or "real", on the other (Niker 2013: 40; Saghai 2013: 488; Rebonato 2012: 203). The thought is that although nudges preserve freedom of choice in a *formal* sense in virtue of not foreclosing options, it is far less clear that they preserve it in a more *substantive* sense. Thaler and Sunstein gesture towards the need for the latter when they state that nudges should be "easy and cheap to avoid"; nonetheless, they also confess that they "do not have a clear definition of 'easily avoided'" (2009: 6, 248-249; cited in Saghai 2013: 488). I suspect that this is due to their awareness that the target of nudge interventions cannot always opt out of the arrangement easily. Given the extent to which these authors are impressed by the empirical evidence about the influence that seemingly insignificant situational factors can have on a person's perception of the choice environment and resultant behaviour, it would seem strange for them not to be

[35] Kelly, for instance, has presented two systems on which we might theoretically base the design of nonpaternalistic nudges – "utilitarian nudges" and "Rawlsian nudges" – that seek to preserve freedom of choice while attempting to steer people's behaviour towards "morally appealing" choices (2013: 222-225).

cognisant that nudges can operate by steering people's cognitive processes in ways that call his or her real freedom of choice into question.

Some have sought to deal with this problem by seeking to find an adequate *resistibility* criterion for nudges, such that an influence cannot count as a "nudge" unless it meets this condition for protecting freedom of choice (e.g., Saghai 2013). This is part of a wider discussion in the literature relating to the supposed difficulty over defining what does and does not count as an instance of nudging. Like Moles, though, I am "not too optimistic about the usefulness of working out a satisfactory definition of nudge" (2015: 648). The political morality of nudging rests not on semantic issues, but rather on substantive normative arguments. In my view, nudging should be understood as a broad category of influence that includes "all measures that involve the (re-)design of people's choice architecture on the basis of psychological insights" (Schubert 2015: 6). As a result, "nudging", "choice architecture", and "behavioural interventions" can be used as synonyms, and will be used as such in this thesis.

This acknowledges that the general category of influence highlighted by the psychological evidence is capacious and complex. It is true that, on a spectrum, nudges (based mainly on nonargumentative influence) fall "between reason and coercion" (Blumenthal-Barby 2012). Nonetheless, what will become clearer when we start to develop a more nuanced normative assessment of this middle category – in Chapters 4 and 5, in particular – is that the seemingly sharp lines between nonargumentative influence, on the one hand, and both rational persuasion and coercion, on the other, are in fact blurred in interesting and significant ways. Nominal freedom of choice is not a concept that will do much work in such an analysis. Parsing this normative terrain will require substantive engagement with concepts such as *autonomy*, *disrespect*, *manipulation*, *well-being* and so on when considering what it means for a nudge intervention to be "liberty-preserving" in the appropriate sense.

§2.3: Mapping the Critical Terrain

This last point leads us into a discussion of the moral constraints on public choice architecture, which I will map out briefly by reference to some of the main strands of critical discussion in the literature, before offering an indication of the gap that this thesis's argument intends to fill.

Nudging has been criticised from many different directions. Arguably the main strand of critique, comprised of a family of related views, is offered by those offended

by its claim to liberal credentials. One such view disputes the grounds for a move towards paternalism by challenging the first part of Thaler and Sunstein's tripartite argument. The most compelling case against paternalistic intervention is not reliant on people being the best judge of what is in their own interests, as the pro-nudge argument contends. Rather, it is grounded in the interest that citizens have in leading their own lives and, as a result, on the centrality of the value of autonomy for liberalism. So, while we can agree with Thaler and Sunstein that the empirical evidence causes problems for anyone who bases their anti-paternalist argument on the best judge assumption, this does not harm anti-paternalism based on autonomy.[36] This raises the worry that nudge paternalism is objectionable on account of the disrespect it shows for people's autonomy.

This concern with autonomy also highlights another issue, one that relates to the manner in which nudges operate when they influence citizens' behaviour. Nudges are the result of a policy design process that is heavily informed by knowledge of our cognitive biases, motivational shortfalls, and other causes of predictable irrationality; and, although it is this feature that makes it an innovative and potentially very effective policy tool, this operational nature makes it vulnerable to several normative concerns. The worry that nudge interventions work by exploiting people's behavioural and decisional flaws and foibles in order to promote certain pre-selected outcomes is at the heart of three related but analytically distinct objections relating to transparency, the harnessing of nonrational tendencies in a way that blocks reasons-responsiveness, and manipulation (see, e.g., Hausman and Welch 2010; Wilkinson 2013; Hansen and Jespersen 2013; White 2013). What tends to unite these three objections is a concern about interference with the autonomy of those being nudged. This seems to be the case whether nudging (i) occurs without the target of the intervention being aware they are being nudged, (ii) works by harnessing automatic flaws in our thinking, or (iii) acts on us in a way that deliberately bypasses or subverts our rational capacities (Dworkin 2017).

This is related to a further concern. Although nudges can be effective at bringing about behavioural change, there is a worry that they are unable to bring about the more substantive shifts in people's beliefs and value structures that are needed to bring about more lasting behavioural change. They generate context-dependent changes in behaviour, without having an impact on the quality of decision-making. Indeed, some worry that this extends beyond the charge of being non-transformative by contending

[36] See Husak (1980: 27-28) for a dismissal of the chances of formulating what he calls "utilitarian objections to paternalism".

that the likely effect of using this policy tool over time would be that citizens' rational capacities may atrophy. This issue shares a similar character with those outlined above: it challenges the strict focus on outcomes by pointing to important process-based aspects of permissible and sustainable behaviour change policies. As Moles highlights:

> "There are two different ways in which people's choices can be improved. First, a person can make a better choice by improving the process by which she reaches her decision. […] The second way of improving someone's choice is to make sure that her decision aligns with her preferences. This case pays no attention to the process of decision-making: it looks only at whether people's choices correspond to their preferences. Nudging is controversial because it employs the second strategy. It enrolls (or exploits) cognitive biases in order to improve choice, without enhancing the process of decision-making." (Moles 2015: 647)

In light of these worries (or any particular set of them), many liberal critics have argued that we should reject nudging (Goodwin 2012), and some have set out alternative proposals for how we should respond politically to the psychological evidence. As one might expect, these are focused on the promotion of personal rather than merely behavioural change through engagement in deliberative democratic fora (John et al. 2011) and certain educational reforms (Gigerenzer 2014).[37]

Nudging has also been criticised from the other side, that is, for not being paternalistic enough. This second strand of critical reaction comes from those who believe that the empirical evidence supports, and sometimes even morally necessitates, a much bolder welfare-protecting response from governments. The main proponent of this position is Sarah Conly, who argues that the correct response to the behavioural science evidence is not to endorse nudging, nor to persuade or educate, since neither option is sufficiently effective at promoting welfare; rather, the correct response is "to save people from themselves by making certain courses of action illegal" (2013: 1). The moral methodology grounding Conly's analysis is robustly consequentialist. She admits that, ideally, the best way to save people from the results of cognitive error would be through the type of public education traditionally advocated by liberalism. The problem, however, is that in the real world this is an ineffective way of helping citizens to make good decisions, and so is not supported by a cost-benefit analysis. Conly offers a forceful critique of nudging, criticising it for giving us "the worst of both worlds" (2013: 8). Interestingly, she agrees with liberal critics that it operates in a manipulative fashion

[37] I explain and engage with these alternative proposals in more detail later in the thesis (§6.2.3).

and claims that this fails to render it any more palatable to the liberal than coercion; and, at the same time, she claims that it fails to "give us the results we want, because people still have the options to pursue bad courses of action", something that is ruled out by its "more intrusive" coercive counterpart (2013: 30-32).[38]

Over the course of the thesis I will show that both sets of critique capture something important. What we need to recognise at this point, I contend, is that our normative analysis is not going to get very far if it addresses nudging in a unitary fashion, by taking a wholesale approach to either rejecting or defending its use in behavioural policy. It is a broad category comprised of various different mechanisms of influence, and it might be used to achieve a variety of different ends. An account of its overall permissibility will need to get clearer on both of these factors and, more specifically, on the interactions between them, since what counts as a legitimate means by which to deliver a particular policy sometimes or often depends on the end that is being served. With respect to thinking about nudging as a means, we need to delineate a number of relevant questions. Does the nudge direct the individual towards making a decision or making a particular choice? If it directs her towards making a particular choice, does the nudge get the individual to respond to the reasons that apply to her? Does the nudge engage an individual's deliberative processes or operate via her nonconscious cognitive processes? Is the influence endorsed by its target in some relevant way? There are a similar set of questions with regards to ends. For instance, is the nudge aiming primarily to induce reasoning or to bring about some pre-selected behavioural outcome? Is it directed towards changing the individual's self-regarding behaviour or her other-regarding behaviour? If it aims to change her other-regarding behaviour, is this for the sake of increasing social welfare or for the sake of moral improvement? If the latter, might this be pursued with the aim not only of preventing others from being harmed or wronged, but also promoting moral insight, responsibility, or virtue in the target herself; and, does this count as supporting her in living well?

It is only by getting to grips with the ethical, moral, and political complexities of this conceptual and normative terrain that we can offer a complete answer to the thesis's main question. In doing so, the answer defended in these pages aims to take account of the problems with nudging while still preserving its promise. Many liberals

[38] Two small points on this argument. First, the manipulative quality of nudging is not considered morally wrong on this view; but, given that it works in this way, there is no gain (and in fact there is likely to be a loss) in leaving open the options to pursue bad courses of action, according to a cost-benefit analysis. Second, on my view the assumption that coercive measures are "more intrusive" than nudge interventions is not necessarily true. This depends on what type of nudge is at work.

have rejected choice architecture as a permissible policy tool; but I will argue that a more nuanced analysis allows us to cast to one side those forms of choice architecture that are objectionably manipulative or paternalistic, and to highlight and move forward with only those forms that are able to support and enhance our agency in various ways. Contrary to Moles's description quoted above, I will argue that nudging can improve people's choices in the *first* sense, that is, some nudges can play a role in improving the process by which an individual reaches her decisions. Despite the focus on individual welfare-promotion and paternalism in the literature, this agency-enhancement need not be confined to self-regarding behaviour; in fact, much of the promise of choice architecture, in my view, comes from its ability to facilitate moral compliance. This is something that has received almost no attention in the literature, but is surely an integral part of a complete assessment of nudging's moral permissibility.[39]

§2.4: Conclusion

The means of influence that this thesis is investigating is grounded in a new model of human agency. This chapter began by outlining the psychological evidence relating to dual-process cognition and the impact that this has on our judgements and behaviour under certain conditions. It moved on to discussing the general move to thinking about these findings as relevant to policymaking, and to laying out the main argument that has been offered as a justification of the governmental use of psychologically-informed policy tools. This argument in favour of nudge paternalism was shown to be imprecise in various ways and the discussion of these points served to clarify the character of nudging, as it will be understood in this thesis at least. This is one of two ways in which this chapter has laid some of the foundations for the argument to come. The second came in the form of setting out the contours of the current debate over the ethics of nudging and, more particularly, in providing an insight into how the type of account I will defend is both related to and distinct from the positions in the existing debate.

[39] Although not as she intended, this strand of analysis engages with the spirit of Conly's critique and sets out conditions for when it would be morally desirable to use coercive measures rather than nudges (§5.3).

3. Nudging Citizens towards Living Well

Our project is to address the question of how, as a means of improving citizens' lives, governments might permissibly apply the recent psychological evidence to the design of public policies and social institutions, if at all. Central to our inquiry, then, is the need for some standard for determining what it is that makes a person's life go well; a standard, that is, of what it might mean for public choice architecture (hereafter 'PCA') to be "welfare-promoting". This chapter outlines a (partial) conception of what it is for a person's life to go well, for the person whose life it is, as being the appropriate one for helping us to set out a satisfactory account of the ethics of public nudging. In so doing, it clarifies where this approach departs from that taken by Thaler and Sunstein (2009) and Sunstein (2014a, 2016a) – a departure that has important implications for the normative analysis of nudging as a policy lever, as we will see in later chapters.

§3.1: Making Citizens "Better Off": Two Views

PCA is commonly justified on account of its potential non-coercively to steer people's choices in ways that will make them "better off" (Thaler and Sunstein 2009: 6). To put it another way, nudge advocates argue that liberal states should nudge citizens, and that they should do so for the purpose of improving those citizens' lives. The extended argument that I present in this thesis agrees with these claims at the *general* level: under certain conditions, it is morally permissible for governments to design PCA so as to improve citizens' lives, and indeed states have good reasons to do so. However, the *particular* view of permissible nudging that I defend diverges from Thaler and Sunstein's in important ways. Our views part ways on account of the different approaches we take to defending the general view. Part of this disagreement rests on differences in our accounts of what it means for a person's life to go well and, as a result, what it might mean for choice architecture to make her "better off". Accordingly, the first section of this chapter maps out this disagreement by outlining (and where necessary, reconstructing) the account of well-being at the base of the standard pro-nudge argument (§3.1.1), before setting out the key elements of my own view (§3.1.2). The next section (§3.2) follows on from this by addressing certain issues relating to why we should favour the latter approach over the former.

§3.1.1: Satisfying (Informed) Preferences

Sunstein does not say much, explicitly, about the conception of well-being that his approach adopts. He offers the following stipulative definition, which he admits is very broad: "With respect to the chooser, let us understand the term [i.e., welfare] to refer to whatever choosers think would make their lives go well" (Sunstein 2014a: 72-73). The central thought is the simple and attractive one that interventions should seek to improve a person's life according to her own standards of what makes for a good life (we might call this the *internalist intuition*). This is generally understood in terms of the satisfaction of an individual's preferences or desires, since people tend to form such preferences and desires for the things that they think would make their lives go well; and there are good reasons to think that this is how Sunstein understands it.[40] The guiding rationale of *preference satisfaction accounts* is encapsulated in Harsanyi's principle of "Preference Autonomy", namely, "[t]he principle that, in deciding what is good and what is bad for a given individual, the ultimate criterion can only be his own wants and his own preferences" (Harsanyi 1982: 55; cited in Scanlon 1991: 24). This refers to a common division between *subjective* accounts, which make well-being depend (solely) upon a person's own desires, and *objective* accounts that make well-being (partially) independent of a person's own tastes and preferences (Griffin 1986: 32).

The motivation for adopting a subjective welfare standard is easy to see. In particular, it gives maximum recognition to the sovereignty of people's individualised preferences (Scanlon 1975: 656-658) – and an ability to respect, and help better to satisfy, people's own preferences is the key justification of this new form of paternalistic intervention. According to Sunstein, choice architecture presents itself as a way – and indeed the best way from within a liberal framework – of "correcting mistakes that people make in choosing the means to satisfy their own ends" (2014a: 164). This highlights a key aspect of the nudge paternalist project as Sunstein understands it: it represents a form of *means paternalism*, because nudges seek not to influence a person's ends (as is the case with *ends paternalism*), but rather try to influence positively her choice of the means so as to help her achieve her self-selected ends. The primary goal of nudging, then, is to design PCA in ways that make it more likely that people "*will promote their own ends, as they themselves understand them*" (ibid.: 19; emphasis in original).

[40] For one, attached to his broad definition is a note referring to Conly's "instructive discussion" of this issue (Sunstein 2014a: 179), in which she states that: "The standard of welfare is […] what is typically termed a subjective one – based on the desires of the subject" (Conly 2013: 102-103).

The key issue with preference satisfaction accounts relates to which kinds of preferences are taken to promote a person's well-being when they are satisfied. The simplest version focuses on the satisfaction of a person's actual desires.[41] Historically, this type of view is linked with the emergence of welfare economics, which rests on the idea that, unlike mental states, a person's preferences can be observed and measured – at least if we rely on a revealed preference framework which posits that a person's actual preferences are revealed in the behavioural choices she makes (Crisp 2013).[42] But the actual preference satisfaction account, even in its most plausible form, faces an "overwhelming" problem (Griffin 1986: 10).[43] Such an account could only be plausible if the preferences are not mistaken, irrational or malleable (i.e., "endogenous"); yet, we recognise numerous ways in which our preferences, as revealed by our behaviour, are affected in these ways. Both Elster (1985) and Sen (1987) have discussed the problem of "adaptive" preferences for preference satisfaction accounts. Indeed, it is additional empirical findings of this kind – such as the evidence of cognitive biases that make us susceptible to behaving in predictably irrational ways on the basis of so-called "welfare-irrelevant" variables – that motivates the nudge paternalism programme in the first place.

For those wanting to insist on keeping with the preference satisfaction account, there is a natural way of responding to this problem. It involves revising the account so that a person's well-being consists in the satisfaction of only those desires and preferences she would have if she were (fully) informed or (fully) rational. Although this move became even more appealing in light of the more recent findings, it had already been made by several economists and philosophers. Harsanyi, for instance, describes the modification thus: rather than taking into account a person's "manifest" preferences, i.e., those manifested in his behaviour, we should take into account his "true" preferences, i.e., "the preferences he *would* have if he had all the relevant factual information, always reasoned with the greatest possible care, and were in a state of mind most conducive to rational choice" (1982: 55; see also Harsanyi 1977: 29-30). Although this conception is not explicitly endorsed, there are a number of references that make it

[41] This can take unrestricted and restricted forms, and it seems clear that the latter are more plausible. This is because unrestricted versions include all the person's actual desires in the welfare standard, which makes for an "implausibly broad range" that would be forced to include, for instance, a person's desire for there to be conscious life of Mars – and it is difficult to accept that the satisfaction of this desire would enhance the quality of this individual's life (Scanlon 1998: 114).

[42] For a critical review of "revealed preference welfarism", see Sugden (1993: 1948-1951).

[43] Nothing hangs on claims about what is the most plausible version, since all versions face the serious charge that is levelled against it here. Having said this, it seems clear that the most plausible account is a *global* version of the *comprehensive* preference satisfaction theory (see Parfit 1984 and Crisp 2013).

clear that this is the account of welfare at work in the justification of nudge paternalism. For instance, Thaler and Sunstein write that:

> "…in many cases, individuals make pretty bad decisions – decisions they would not have made if they had paid full attention and possessed complete information, unlimited cognitive abilities, and complete self-control." (Thaler and Sunstein 2009: 6-7; also see Sunstein and Thaler 2005: 177).

In order to capture the variance between a person's actual preferences and her informed or rational ones, Thaler and Sunstein advance a distinction between "Humans" and "Econs", the latter being an imaginary species modelled on a fully rational, utility-maximising agent not susceptible to bounded rationality, cognitive biases, motivational shortfalls, or preference malleability (2009: 7-9). One way of thinking about the aim of choice architecture, which seems to follow from this, is as a policy tool that seeks to make it more likely that Humans' choices will better track the judgements that Econs would make if they were in their place.[44] In a similar if not somewhat more measured manner, Sunstein contends in his most recent work that, in applying the "as judged by themselves" standard (i.e., the liberty-preserving aspect of nudge paternalism), "choice architects should be interested in choosers' informed judgments" only, and that "it is fair for choice architects to insist that actual judgments do not reflect informed choices" (2016a: 45). Indeed, elsewhere he has called this version of the preference satisfaction view, which includes such an "information requirement on desires" (Qizilbash 2012: 653), the "*informed-chooser approach*" (Sunstein 2015: 73).

Hence, despite the claim that they are "not attempting to say anything controversial about welfare" (Sunstein and Thaler 2003: 1163), it would appear that there is a good case to be made that these advocates of nudging implicitly adopt, and seek to justify their "welfare-promoting" interventions by reference to, an *informed preference satisfaction account* of well-being (see Qizilbash 2012: 651-652, Sugden 2008: 232, Rebonato 2013: 193, and Hédoin 2016 for similar assessments). This adherence to the

[44] Indeed, Sunstein himself acknowledges as much when he quotes Rebonato's characterisation of nudge paternalism: "Libertarian paternalism is the set of interventions aimed at overcoming the unavoidable biases and decisional inadequacies of an individual by exploiting [Sunstein suggests "counteracting" instead] them in such a way as to influence her decisions (in an easily reversible manner) towards choices that she herself would make if she had at her disposal unlimited time and information, and the analytic abilities of a rational decision maker (more precisely, of *Homo Economicus*)" (Rebonato 2012: 84; cited in Sunstein 2013b: 1860). Interestingly, although Sunstein says that this "helpful" definition is nonetheless "imprecise" in various ways (one of these is noted above), he does not reject or seek to modify the claim that nudge paternalism understands the welfare goal in terms of what a rational decision maker (*homo economicus*) would choose (Hédoin 2016).

preference satisfaction account of well-being has problematic implications for their views on permissible paternalistic intervention, which we will explore below (§3.2.2). For now though, we can call attention to a fundamental tension in the standard nudge policy programme. On the one hand, it wants to take seriously people's assessment of their own good; the "as judged by themselves" standard is designed for this purpose. On the other hand, the initial justification for nudging is that people's actual preferences and judgements do not track their welfare, and so the welfare-promoting aim of these interventions requires choice architects to steer people in the "right direction" (Sunstein 2014a: 17), namely, the direction they *would have* judged for themselves if they were sufficiently or ideally informed or rational. This has been called the "Soft Paternalist's Paradox" (Grüne-Yanoff 2009). Since this second part seems to involve sacrificing the values that the first part is supposed to protect, serious questions are raised about the ability of informed preference satisfaction views to deal with the issues relating to personal autonomy that are at the core of discussions over (permissible) paternalism.

§3.1.2: Autonomy and Living Well

It is this point that my view seeks to push back on. I do not offer a full account of the good life for the person whose life it is. It may be, as others have argued, that "prudential deliberation" would direct us towards accepting that a good life for humans would include certain elements or "prudential values", such as pleasure, deep personal relationships, the accomplishment of goals we take to be important, and so on (see, for instance, Griffin 1998: 29-30); but I do not take any view on these more expansive issues here. Rather, I offer only a partial account of well-being that claims that personal autonomy has value for a person's life that goes beyond the value that it derives from other things which it makes possible, such as, for instance, the satisfaction of a person's preferences or desires. The claim that autonomy has this intrinsic, and not only instrumental, value for a person's life thereby makes this partial account of the human good *objective* in nature, understood in the minimal sense outlined above.[45]

The main motivation for this sort of view is the simple thought, and one which garners broad support, that a central part of what it means for a person's life to go well for her is that she lives it in an active and authentic way, a way that is shaped by her own sense of what makes for a good life. There are different ways of capturing and expressing this broad ethical ideal. Ronald Dworkin, for instance, maintains that, "We

[45] For some, Sunstein included, this objectivity makes this sort of standard a controversial one. These issues are discussed in the next section. Here, I focus on presenting the key elements of my (partial) view.

cannot explain the importance of a good life except by noticing how creating a good life contributes to living well" (2011: 196). As part of his ambitious interpretative project in *Justice for Hedgehogs*, Dworkin argues that living well requires "that you live in response to […] your situation and the values you find appropriate", and that doing so requires ethical independence in our relations with others, since: "Authenticity is damaged when a person is made to accept someone else's judgment in place of his own about the values or goals his life should display" (2011: 210-212). Dworkin also clearly states the objective nature of this ideal: "we think that authenticity is not a taste but a necessary virtue, that there is something *wrong* with an inauthentic life. We think that authenticity has objective importance" (2011: 213; emphasis in original). Joseph Raz also contends that autonomy is an ethical ideal related to the creation of one's own life, and that an autonomous life is marked "not by what there is in it but by how it came to be" (Raz 1986: 371). Like Dworkin, he describes it as in part a social ideal which "designates one aspect of the proper relations between people" (ibid.: 378). He writes that the ideal of autonomy, as a conception of well-being, "transcends the conceptual point that personal well-being is partly determined by success in willingly endorsed pursuits and holds the free choice of goals and relations as an essential ingredient of individual well-being" (1986: 370). James Griffin also includes autonomy within his category of "the components of human existence", which he marks out as one of the core values in his prudential list view. He justifies this in the following way:

> "Choosing one's own course through life, making something of it according to one's own lights, is at the heart of what it is to lead a human existence. And we value what makes life human over and above what makes it happy. What makes life 'human' in the special normative sense that the word has here centres on 'agency'. One component of agency is deciding for oneself. Even if I constantly made a mess of my life, even if you could do better if you took charge, I would not let you do it. Autonomy has a value of its own." (Griffin 1998: 29-30)

In a similar fashion, John Stuart Mill grounds his liberal theory in the claim that the "free development of individuality is one of the leading essentials in well-being" (2005: 69). Although it is possible that we can be "guided in some good path" in a way that does not make use of our distinctively "human faculties of perception, judgement, discriminative feeling, mental activity, and even moral preference", the value of such an action is greatly reduced because it was not done autonomously. For Mill, it "really is of

importance, not only what men do, but also what manner of men they are that do it" (2005: 71-72).

There are numerous conceptualisations of personal autonomy within the literature. I leave the fuller presentation of my conception until the next chapter (see §4.2.1). Here, I set out two features of the autonomy-based account of well-being that characterise my approach to assessing the permissibility of nudge paternalism.

The first thing to note is that, in the prudential realm, the view of autonomy that I defend is *procedural* or *content-neutral*, as opposed to substantive. Such accounts deem a particular action autonomous if it meets certain procedural criteria, and thereby remain neutral with respect to what a person conceives of as worthy of pursuit. As Christman explains, "there is no requirement that the autonomous person must pursue the good, except in the sense that she pursue things that mesh with her ongoing evaluative orientation (so in that sense she must aim at what she herself takes as minimally good)" (2017: 13). Autonomy both obtains and has value for a person's life so long as certain authenticity conditions and competency conditions are met: generally speaking, the former typically relate to a person's capacity to reflect on and appropriately respond to her desires, values, and so on, while the latter specify that she must have various capacities for practical reasoning, self-control, etc., and that she be free to exercise these critical capacities without undue internal or external interference (Christman and Anderson 2005: 3). To be ourselves is to be committed to a set of pro-attitudes that create reasons for us to act in particular ways; and the activity of responding to these reasons, and thereby responding to our own reasoning about what is a good life for us, has significant value for our lives.

This explains why the evidence from the cognitive and behavioural sciences is troubling from this perspective on living well. It is not problematic, primarily, because a person's choices are less likely to produce good (i.e., "welfare-promoting") results, as Sunstein contends. On that view, the promise of PCA comes from its ability to change the choice environment so that the behaviours that result from automatic and context-dependent cognitive processes are more aligned with how a person would have chosen to act, if she were more like an Econ. Instead, these cognitive processes, and the behaviour they generate, are troubling in so far as they pose a threat to our ability to live active, authentic lives. On this view, choice architecture holds out a promise, but it also introduces an additional concern. Threats to autonomous agency can be internal, as well as external; and, in some cases, automaticity and context-dependency render us passive,

in an important sense, with respect to our ability to respond to our own reasoning about how to act in certain situations. The promise, then, is that there might be ways of designing PCA that help citizens to overcome this internal threat to autonomous agency, thereby promoting this important constituent of human well-being. But the nature and value of autonomy places important constraints on the permissibility of nudge paternalism: using choice architecture in ways that exploit what we know about our cognitive processes in order to generate particular behavioural outcomes, for example, may threaten both the authenticity and competency conditions of autonomy (more on this in §4.2.2).

A second point to clarify is how the value of autonomy, understood as a constituent of living well, relates to morality. The relationship between the human good ("living well") and morality ("living right") is a complex issue, and one of the central questions in moral philosophy (Nagel 1986, Ch. 10). Philosophers have set out a number of different theories about this relationship (see, e.g., Crisp and Hooker 2000; Griffin 1998). Broadly speaking, I side with those philosophers, such as Aristotle (2009) and Dworkin (2011), who view "ethics and morality as deeply complementary" (Jolls 2010: 641), rather than as conflicting with each other in a deep way.[46] Specifically, I think that an individual's own well-being can be served by complying with her moral reasons, where this is distinguished from mere conformity by the fact that it requires performing the morally required or desirable action in direct response to the reason (Raz 2000: 216; for more this distinction and for a full account of this view, see §5.2.2). When an individual treats others as she ought to treat them, there is ethical value for that individual, on my view, when she does this autonomously – as a response to her own reasoning about how to act in that situation, rather than, e.g., because she has been forced into conforming behaviour. This is not the same as the value of acting autonomously in the prudential realm, because its value is not content-independent.

We can distinguish between two broad views on the value of autonomy: the *unconditional* view and the *conditional* view. The former claims that personal autonomy is always *pro tanto* good for the person (even if it is ultimately outweighed by other reasons). By contrast, the conditional view contends that autonomy is not valuable to a person's life when it is used in selecting an immoral option.[47] There are two possible

[46] Recall that the term *ethics* is used to designate how we should live if we want to live well (Dworkin 2011).

[47] Some also include "worthless" options in this category alongside morally repugnant ones (e.g., Raz 1986). In my view, Raz problematically runs together ethically (or prudentially) worthless and morally

versions of this view. The first, stronger version claims that when an individual autonomously acts in an immoral way, this is *worse* than if she did so nonautonomously. Raz, for instance, takes this route when he claims that the "wrongdoing casts a darker shadow on its perpetrator if it is autonomously done by him" (1986: 380). The second version claims that when a person autonomously chooses an immoral option, this *cancels out* the value that would have otherwise attached to acting autonomously (Dworkin 1990). My account subscribes to the conditional view on autonomy's value, though I remain agnostic about which of its two versions better tracks the truth. In the moral realm, then, what counts towards living well is *morally responsible agency*, i.e., autonomous agency constrained by a more stringent reasons-responsiveness requirement.

Clarifying these two features has helped to see the general shape of the autonomy-based approach that I take towards assessing the permissibility conditions for nudging, when done for the sake of promoting the target's well-being. Some might question whether taking this approach, as compared to Thaler and Sunstein's, would make much of a difference to the overall normative assessment of PCA. As we will see in the remainder of the thesis, this worry is shown to be groundless, because it does have significant implications for the analysis. Thaler and Sunstein (i) accept only a very "thin" sense of autonomy, essentially equivalent to having the freedom to choose, as having relevance for the ethical assessment of nudging, and (ii) do not count (even this conception of) autonomy as an intrinsically valuable consideration. On their view, the value of freedom of choice is "doubly contingent", since it depends both on whether a person has a preference to make a choice themselves (if so, then freedom of choice is "an ingredient in welfare", as the act of choosing would count in the welfare calculus[48]), and on whether her use of the freedom to choose actually does promote her welfare, as these authors understand it (Thaler and Sunstein 2003: 1198).

There are clear points at which our analyses will diverge. For example, the welfarist approach can care that people's informed preferences are satisfied, but it cannot care about whether or not people deliberate and reason well in coming to that outcome, unless a person's preference for autonomy is strong enough to outweigh the welfare gain that is taken to be attached to the particular outcome. In short, it separates the outcome from the process of getting to that outcome, and places value nearly

repugnant options in his discussion, thereby making his view too substantive in the prudential realm (see Christman 2017 for a fuller discussion of this issue).

[48] Sunstein writes: "Perhaps welfare is what really matters, and perhaps autonomy is important […] not because it is a genuinely independent value but because people sometimes become frustrated and angry if they cannot get their own way – a point about welfare" (2014a: 22).

exclusively on the former. As a result, it does not have any concern for whether an individual deliberates about her actions, acts in ways she accepts as acting on the basis of her own reasons, responds sensitively to moral reasons, and so on, so long as her welfare (that is, the satisfaction of the preferences she would have if she were an Econ) is promoted. An account of well-being that includes autonomy as a central component, by contrast, holds that a person's well-being is advanced by measures that support her in attaining the competency and authenticity conditions for autonomy. It recognises that our autonomous decision can be more or less welfare-promoting in the broader sense, but it is committed to the view that there is ethical value to the practical reasoning and decision-making processes involved in autonomous and morally responsible agency. And, just to be clear, this view is not committed to the implausible view that personal autonomy is all that matters. It is a partial account of living well, which recognises that autonomy is one, but only one, component of human well-being and that, as a result, "sometimes reasons to promote autonomy should give way to reasons to promote other ideals" (Wall 1998: 185).

§3.2: Autonomy, Perfectionism, and Anti-Perfectionism

Moral theories can recognise the value of autonomy in different ways. By basing this on a partial account of living well, the thesis's approach to assessing the permissibility of nudge paternalism admittedly rests on a potentially controversial claim about well-being. This raises questions about the reasons that move states to act. This section specifies in more detail the theoretical premises of my approach and how these relate to alternative views (§3.2.1), and offers some support for my approach by highlighting a series of problems that bear on the consistency of Sunstein's alternative approach (§3.2.2).

§3.2.1: Autonomy, Advantage, and Moderate Liberal Perfectionism

Some people worry about the state acting on some types of valid reasons when justifying political action. One way of mapping out this issue involves distinguishing between two ways in which the question 'What does human well-being consist in?' can be interpreted and answered. In the first instance, when we ask this question we often mean to ask 'What fundamentally makes one's life go well?' The term "well-being" is commonly used in philosophy to describe what is "non-instrumentally or ultimately good *for* a person" (Crisp 2013; emphasis in original). As such, it attaches pre-eminently to the lives of individuals and is understood as being more or less the same as his or her

interest or (non-moral) good. A second approach, by contrast, means to ask: 'What should we take as our metric for making intrapersonal and interpersonal comparisons of well-being for the purposes of making distributive and policy decisions?' If our aim is to reduce inequality, to give priority in decision-making to the worst off, or to attend to the claims of an individual, for example, it is necessary that we have some appropriate way of comparing the situations of individuals, so that we can identify the advantage of different individuals and how much more or less they would be advantaged by a particular policy. To capture this difference, we might call the first the *well-being question* and the second the *advantage-for-policy question*.

In these terms, we can see that what is needed with respect to our assessment of nudge policies is an answer to the advantage-for-policy question. Nevertheless, there is also an important issue about how one understands the relationship between these two questions. It might seem natural, for instance, to think that the criteria for making policy decisions is provided by one's answer to the well-being question; and indeed some do take this view. But others believe that when identifying levels of advantage for the purpose of making policy, it would be wrong simply to adopt the best conception of what makes an individual's life go well. A common basis for why this might be considered wrong is that it would be a disrespectful response to ethical pluralism, since it is a fact that, in heterogeneous societies like ours, people disagree about whether hedonism, honouring a particular deity, or success in various kinds of valuable relationships and projects, is constitutive of one's life going well. The central claim is that if our metric for identifying advantage for policy were to rest on a controversial view of well-being, such as one of these, it would be rejected by some or perhaps even many citizens (Rawls 1993).

This disagreement is at the heart of the debate between the two camps of theorists referred to as perfectionists and anti-perfectionists. Broadly speaking, theorists in the first camp "advance an objective account of the good and then develop an account of ethics and/or politics that is informed by this account of the good" (Wall 2012). In the terms introduced above, perfectionists rely on (some part of) their answer to the well-being question to direct, to some extent at least, their conception of advantage for political purposes. Anti-perfectionists, by contrast, maintain that the state should not promote ways of life that are valuable; rather, it should adhere to the *principle of state neutrality* when making policy. There are different ways of formulating this neutrality constraint, though arguably the most forceful is when it is done by appeal to

the ideal of public reason. Here, the key thought is that the state ought to justify the exercise of political power by appeal only to reasons that all free and equal individuals "may reasonably be expected to endorse in the light of principles and ideals acceptable to their common human reason" (Rawls 1993: 137; see also Nagel 1991). This requirement for shareable reasons excludes appeal to controversial ideals of the good, and instructs that we bracket our full understanding of the human good when advocating for and justifying political action.[49] As a result, it requires that we answer the advantage-for-policy question in a way that not only does not rely on, but is also neutral between, any (reasonable) answer to the well-being question.[50]

As already noted, according to these common formulations, the theory of the political morality of nudging I defend is a liberal perfectionist one. It assumes a conception of advantage for assessing nudge interventions that is based on a claim about the objective value of leading an autonomous life for living well. It is beyond the scope of the thesis to defend this claim in detail and so, in this respect, it operates more in the role of a premise for my extended argument, rather than as the conclusion of any argument that has been provided.[51] But it is worth making some comments about the character of this claim that lessen its controversiality. In particular, this claim represents a *moderate liberal perfectionism*. It is moderate for a number of reasons: first, it promotes only a certain category of goods, namely, *agency goods*, and so supports particular ways of life only in so far as they involve these goods relative to others which do not; second, it uses *noncoercive* means, and even its use of these noncoercive nudge mechanisms is restricted to those able to create a social environment that is conducive to the promotion of valuable agency goods; and third, it insists that the pursuit of these agency goods is "tempered by other values", rather than being considered the only intrinsic value, thereby making it a form of what Chan calls *mixed* perfectionism (Chan 2000: 14-

[49] Wall refers to this version of anti-perfectionist political morality as the "bracketing strategy" (1998; see Ch. 3). One of the main problems with this strategy is raised by the *asymmetry objection*. This objection highlights what some take to be a "serious internal challenge" to the anti-perfectionist position by asking: "Why does the state's pursuit of controversial conceptions of the good life *alone*, but not of controversial conceptions of other matters [e.g., social justice, criminal justice, education, and national defence], lead to instability and illegitimacy?" (Chan 2000: 9). This same objection has been set out in numerous ways. For some of these, see Mulhall and Swift (1992: 234); Sandel (1994: 1782-1789); Waldron (1994); Caney (1995: 257-258); Chan (2000: 20-42); and Fowler and Stemplowska (2015). For attempts to respond to the asymmetry objection, see Clayton (2006: 19-27) and Quong (2010, Ch. 7).

[50] For Rawls (1999), this requires a particular "resourcist" answer based on "primary goods", which are goods that are supposed to be desirable for every human being and useful in their pursuit of their conception of the good.

[51] Pursuing this type of argument would be the subject of a thesis in itself. A number of theorists have offered book-length defences of liberal perfectionist accounts of political morality. For example, see Raz (1986); Wall (1998); Sher (1997); Kramer (2017).

15). This moderate perfectionism, I argue, is a natural response for those who value autonomy to the evidence that the liberal state is unavoidably involved in structuring the landscape of choice. In other words, I argue that – knowing this – it should do so in a way that creates and sustains the conditions that enable people to live autonomous lives, while remaining ecumenical on specific matters relating to prudential goods.[52]

It is worth noting that my approach is not necessarily perfectionist on account of claiming that (1) a liberal state ought to nudge, (2) when it can do so for the sake of promoting autonomous agency, (3) since if it does not do so, it leaves citizens subject to unnecessary disadvantage. What makes it perfectionist is the way in which I defend (1)-(3) by tying this set of claims to living well. It might be possible for this set of claims also to be defended by tying them to something else, such as citizens' interests in exercising their moral powers, for instance (Rawls 1999). In other words, "Some liberal neutralists might argue that they can justify the policies favoured by perfectionists by nonperfectionist arguments" (Chan 2000: 19). Indeed, there has been a renewed interest in recent years in (re)considering the relationship between anti-perfectionism and autonomy – or, more precisely, the state's role in promoting citizens' personal autonomy (Colburn 2010; Christman 2017; Nye 2012; Porter 2011; Quong 2010). Some argue that the view that autonomy is valuable and should be promoted by the state –a view described as "autonomy-minded liberalism" (Colburn 2010) – is the paradigmatic example of a liberal perfectionist political morality and so is necessarily opposed to anti-perfectionism (Quong 2010: 45-72). In contrast, others have argued that there is no necessary tension in liberals holding commitments to both autonomy-mindedness and anti-perfectionism (Colburn 2010: 43-68). It remains an open question, therefore, whether anti-perfectionists would reject the (substance of the) autonomy-based view of permissible nudging that I defend in this thesis.[53] To put it in more positive terms, there is a separate, further question about whether the account I present could be nested within an anti-perfectionist framework.[54] For the purposes of the thesis, however, I

[52] For a full defence of moderate perfectionism, which makes the strong claim that this type of state perfectionism is "desirable, unavoidable, and legitimate", see Chan (2000).

[53] For example, Moles's anti-perfectionist account of permissible nudging defends a weak form of nudge paternalism directed at promoting (a particular Frankfurtian conception of) autonomy, which he grounds in Rawls's idea in *A Theory of Justice* that in the original position "it is rational for (the parties) to protect themselves against their own irrational inclinations by consenting to a scheme of penalties that may give them a sufficient motive to avoid foolish actions and by accepting certain impositions designed to undo the unfortunate consequences of their imprudent behavior" (1999: 219; cited at Moles 2015: 664 and Husak 1980: 39). For the discussion of this "autonomy-friendly" nudge paternalism and its compatibility with Rawlsian political liberalism, see Moles (2015: 664-667).

[54] As Wall (2012) states, although much of the debate over state neutrality "assumes that there is a strict incompatibility between state neutrality and perfectionist politics", the reality is that the relationship

leave this issue to one side and accept the liberal perfectionist foundation of my account.[55]

§3.2.2: Facing Up: Problems with Sunstein's Approach

One way of offering support to my approach is to reveal problems with its main rival in this debate, namely, with the preference satisfaction approach outlined in §3.1.1. To this end, this subsection briefly outlines three kinds of related yet distinct problems, all of which call into question the consistency of Sunstein's approach. Since this alternative approach explicitly expresses the aim of standing opposed to perfectionism, this negative argument serves to provide support for the approach I take (at least vis-à-vis Sunstein's).

Now that we have made the distinction between the two questions above, we can see more clearly that nudge advocates offer a particular, welfarist answer to the advantage-for-policy question. This seeks to justify nudge paternalism along the following lines: (1) a liberal state ought to nudge, (2) when it can promote welfare, understood as the satisfaction of citizens' informed or rational preferences, (3) since not doing so leaves citizens unnecessarily disadvantaged. One of Sunstein's main motivations for holding a preference satisfaction view of advantage comes from an ambition to avoid perfectionism. He is explicit that he is interested in defending nudging only as a form of means paternalism; and the thought is that interventions would not be means paternalist if the welfare standard one adopts is objective, since such interventions would have the aim of directing a person's actions towards certain ends that are taken to be valuable, regardless of whether the person has a preference for them. The "as judged by themselves" (hereafter 'AJBT') standard is the main way in which the welfarist conception of advantage retains the subjectivity required for nudging, as an all-purpose policy tool, to be considered means paternalist. Sunstein does not offer any arguments against perfectionism, but emphasises that in so far as choice architects adopt the AJBT standard, "they reject perfectionism, and they do so on principle" (2016a: 51).[56] The rationale for this principled rejection of perfectionism

---

between them "is much more complex", especially in the case of those views that embrace the fact of ethical pluralism, as mine does.

[55] Parts of my discussions in Chapters 4 and 5 might be viewed as having some bearing on this question, given their ambition to show that some of the worries that anti-perfectionists often have about perfectionist political action are unwarranted in this case.

[56] Although he does make reference to the objection(s) to the AJBT standard that might be offered by some perfectionists (namely, that it is "too subjective"), Sunstein states only that: "These questions raise serious questions within political philosophy, which I cannot adequately answer here" (2016a: 51).

seems to be: public policy ought to be done on the basis of preference satisfaction because we (i.e., advocates of nudging) do not want to gainsay people's views about what makes for a good life, and the best way of avoiding this is to operate with a conception of satisfaction that makes reference to people's own preferences as indicators of whether they are more or less advantaged. This involves making a value judgement, as all conceptions of advantage must, but resists appealing to any particular view about what makes someone's life go well. Let us suppose, then, that we accept this ambition to avoid perfectionism. Would such an ambition lead us to accepting Sunstein's view?

The first issue is grounded in the observation that it is not at all obvious that preference satisfaction is the neutral, subjectivist welfare standard that it is assumed to be. Indeed, it is odd that Sunstein seems to assume its neutrality, especially when the leading liberal theorists who care about avoiding perfectionism reject welfarist views of advantage-for-policy (e.g., Nagel 1986; Dworkin 1990; Rawls 1993). Although the aim of adopting a preference satisfaction view is to avoid gainsaying people's views about what makes for a good life, and thus to eschew controversial claims about well-being, its adoption by the state would, in the words of Dworkin, "necessarily be imposing on everyone its collective judgment of what lives are good and how to live well" (2011: 355). We can see this by noticing that it is possible that some, perhaps several, citizens might judge their own well-being in ways that make no essential reference to the satisfaction of their (rational) preferences at all. Consequently, one might question why the adoption of informed preference satisfaction as the welfare standard should be seen as any less controversial than the adoption of opportunity for autonomous agency as a form of advantage for policy.

Even if this were not so, and we accepted for the sake of argument that this type of welfarism can be neutral in the appropriate sense, there are questions over translating this neutrality into practice. This second issue challenges the extent to which, in practice, Sunstein's view is able to remain true to its initial motivation. According to Rebonato, nudge advocates "find themselves in a singularly difficult position" because they are "faced with an enormous problem of intelligibility of preferences" (2012: 195-196). Sunstein acknowledges himself that the "real danger" for his view is that, when we move away from people's actual preferences to asking what people would want if they were adequately informed, rational, self-controlled, and so on, choice architects will be relying on their own values and beliefs when designing nudges (2016a: 45). At the very

least, they will be relying on their judgements about what it would be rational for someone to do in a particular situation in order to design PCA that steers citizens in the "right direction". This drives a wedge between the informed preference satisfaction welfare standard and the AJBT standard. Indeed, some might argue that it appears to defeat the whole purpose of adopting the AJBT standard in the first place, which seriously calls into question Sunstein's claim to be defending (only) means paternalism.

One might respond to this by saying that nudges leave open the option to opt out if a citizen believes that it has not, in fact, nudged her in the right direction according to her own conception of the good. This is true, at least in principle. But once we delve a little deeper, this defence begins to look shaky. The third issue I wish to highlight, then, relates to the fact that there are different interpretations of what is required for the AJBT standard, and within this set some interpretations are more attractive than others. What counts as *judging* (according to the AJBT standard) in the case of nudges that operate by taking advantage of our automatic cognitive processes in their efforts to steer behaviour? Thaler and Sunstein work on the assumption that, so long as an individual does not opt out, the intervention can be understood as having promoted her welfare by satisfying some informed preference. In other words, the AJBT standard relies on a revealed preference framework. These authors admit as much when they acknowledge that: "Some readers might think that our reliance on behaviour as an indication of welfare is inconsistent with one of our central claims – that choices do not necessarily coincide with welfare" (Thaler and Sunstein 2006: 254).

This raises a serious concern. The inconsistency that they reference comes to fore because some nudges act on us in ways that may undermine the idea that any judgement has been made by the target at all, never mind the kind of judgement that one might expect the AJBT standard to protect. In fact:

> "if the nudges of the libertarian paternalists – such as changing the default option – are effective, and exploit the decisional inertia of the choosers, then it makes little difference that there is a nominal right to opt out. And if their nudge is *very* effective, then having the nominal right to reverse the nudge makes *very* little difference." (Rebonato 2012: 203; emphasis in original)

If the AJBT standard, as Sunstein uses it, can be expanded so far as to include instances of this kind, then this is surely not an attractive interpretation. An attractive version would need to have a more convincing condition of what it takes for an individual to judge for herself what does and does not count as a welfare-promoting influence. One

interpretation of the standard that does not fall foul of the problem of interpersonal intelligibility of preferences is an autonomy-based one. Again, this could take on different particular features; but taking seriously the AJBT standard may commit us to using PCA to promote autonomous agency: for example, "a social planner could try to favour those social conditions that facilitate the formation of an informed and well-considered choice, and then let individuals get on with freely choosing" (Rebonato 2012: 194). If so, then accepting the AJBT standard does not, as Sunstein claims, require that choice architects necessarily reject perfectionism, at least the moderate liberal perfectionism described above.

As a result of these three issues it has become clear that, despite trying to take the ecumenical high ground against accounts like mine, Sunstein's approach is not as ecumenical as he would think. But if his approach cannot lay claim to the neutrality that it hoped it could, then my moderate liberal perfectionist approach does not look to be as controversial as nudge advocates might like to think. There may be grounds, in fact, to claim that the account of permissible nudge paternalism I develop is *more* ecumenical than Sunstein's, given its focus on procedural agency-based features rather than the rational outcomes that his approach focuses on. Regardless of whether this is true or not, this subsection has highlighted three problems that seriously call into question the coherence and neutrality of the informed preference satisfaction approach.

§3.3: Conclusion

This chapter has outlined a partial, autonomy-based conception of what it is for a person's life to go well as being the appropriate standard in setting out an account of permissible public nudging. In doing so, it rejects the (implicit) informed preference satisfaction view adopted by Sunstein, and has argued that – even if it were an attractive view – it suffers from a set of problems that call into question its consistency. These problems have not been adequately acknowledged by its advocates, but they do potentially undercut the whole analysis of the ethics of nudging that follows from adopting this welfare standard. The conceptual and theoretical work carried out in this chapter and the previous one have laid the foundations for the full normative investigation that follows in Chapters 4 and 5, on prudential and moral nudging respectively.

4. Nudging as a Means of Autonomy Support

This chapter examines whether the political community can ever permissibly use choice architecture for the sake of improving citizens' own well-being and, if so, under what conditions. Its principal aim is to defend the view that public nudging, when done for the target's own sake, is morally permissible only if it is both non-manipulative and autonomy-supporting, where this means that it provides an opportunity for people to be better able to act on their own reasons. This selective account – which I call *autonomy-supporting nudge paternalism* – represents a subtle, middle-ground view between advocates and critics of paternalistically-motivated nudging. In doing so, it sets out the specific and limited role that public choice architecture (PCA) permissibly might play in (supporting) citizens' prudential practical reasoning.

The chapter begins by outlining different types of claim that might be made regarding the relationship between autonomy and PCA, which serves both to set out the main views in the literature and to provide a sense of the general shape of the view I will be defending (§4.1). It then offers a more detailed account of the concept of personal autonomy, and explains the two kinds of relevant threat to this ethical ideal – an internal threat from automaticity, and an external threat from manipulation (§4.2). This raises the question of whether some types of PCA can mitigate the former threat while not falling foul of the latter, and thereby act as a form of *autonomy support*. To assess this, a full examination of the charge of manipulation is carried out, which shows that certain nudges are objectionably manipulative and thus presumptively impermissible, but that others do not operate on their target in a psychologically manipulative fashion (§4.3). The chapter then outlines two categories of PCA – *nudge-autonomy* and *auto-nudge* – that are able to play an autonomy-supporting role, and defends these as permissible forms of nudge paternalism (§4.4).

§4.1: Nudging and Autonomy: Three Types of Compatibility Claim

Traditionally, liberals such as J.S. Mill have used the value of autonomy to a human life to ground the anti-paternalism principle. The psychological evidence canvassed in earlier chapters appears, however, to have provided a justification for a means by which to integrate a new form of paternalism into liberal political theory without offending

against this central value.[57] Nudges claim to represent a "liberty-preserving" mode of promoting well-being, which "should be acceptable to those firmly committed to freedom of choice on grounds of […] autonomy" (Sunstein and Thaler 2006: 233). Yet much of the debate over the ethics of nudging has been infused with a condemnatory attitude, grounded in the idea that public nudges interfere with citizens' autonomy in some objectionable way. This first section seeks to clarify the issues at stake in this debate by distinguishing between three possible views that one might take about the relationship between PCA and autonomy. Doing so offers a way of introducing and problematising the main conceptual and normative considerations that the chapter will need to address, as well as making clear at the outset how the view that I will be developing and defending is situated relative to others in the literature.

§4.1.1: Compatibility and Incompatibility Views

A good way of introducing the first two kinds of view is offered by Hausman and Welch's definition of choice architecture, since one view emphasises its first part and the other its second part:

> "Nudges are ways of influencing choice without limiting the choice set or making alternatives appreciably more costly in terms of time, trouble, social sanctions, and so forth. They are called for because of flaws in individual decision-making, and they work by making use of those flaws." (Hausman and Welch 2010: 126)

Let's begin with the first part and with what we might call the *compatibility view*. Advocates have defended nudging as a permissible means of governmental influence according to the normative standard of liberty-preservation. The fundamental claim, as we saw in Chapter 2, is that these choice-influencing interventions preserve citizens' freedom since they do not coercively restrict or alter choice-sets (Thaler and Sunstein 2009: 6). We might agree, at least for the sake of argument, with this claim about the noncoercive nature of nudging.[58] For Sunstein and Thaler, this is enough to ground a

[57] Note that certain forms of paternalism have always been justified on the classic anti-paternalist view. Mill, for instance, allowed that, in cases of involuntariness like his rotten bridge example, paternalistic intervention is justified because it does not represent "any real infringement of [the individual's] liberty" (2005: 117).

[58] Goodwin challenges this view by arguing that the truth of this noncoerciveness claim hinges on Thaler and Sunstein's particular libertarian account of freedom and so does not hold across other (arguably more attractive) conceptions of freedom (2012: 88). Although I make a similar distinction between conceptions of freedom (and the problematic consequences of this for Thaler and Sunstein's arguments) in what follows, I do so principally for the sake of pursuing a different line of critique which causes problems for the compatibility view *even if* we accept that nudges are noncoercive (for a similar approach, see Mills 2015).

version of the compatibility claim: "we think that respect for autonomy is adequately accommodated within the [liberal] aspect of [liberal] paternalism" (2003: 1167 fn. 22). Thaler and Sunstein's "liberal" account is concerned primarily with coerciveness because it understands (and seeks to preserve) freedom in its negative sense, namely, the sense of being "free to choose" (Thaler and Sunstein 2009: 5; Friedman and Friedman 1980). According to this negative conception, freedom consists in "not being prevented from choosing as I do by other men" (Berlin 2002: 131). It is true that coercion invades autonomy; and so, the (stipulated) fact that nudging does not diminish an individual's options means that it does not fall foul of one of the possible ways in which her autonomy might be threatened. On this view, then, the compatibility claim amounts to a statement that choice architecture is compatible with autonomy in virtue of the fact that it preserves citizens' freedom of choice.

But there are a number of issues with this claim, of which I'll briefly outline two for the sake of laying the groundwork for the opposite view. The first has already been discussed in clarifying the character of nudging (§2.2.2). Thaler and Sunstein's claim can be challenged on its own terms by contending that nudges do not even necessarily preserve freedom of choice. To see this, we distinguished between two senses in which nudges might be said to preserve freedom of choice – a basic sense and a substantive sense. Although it is clear that nudges preserve freedom of choice in the basic sense in virtue of not formally foreclosing options, it is far less clear that they do so in the more substantive sense, since some nudges appear to operate by steering agents' cognitive processes in ways that call his or her actual freedom of choice into question. Leading on from this, the second issue challenges the account of autonomy adopted by Thaler and Sunstein in a more wholesale fashion. Autonomy is generally viewed as equivalent, or certainly very similar, to Berlin's positive conception of freedom (2002: 131-134); not, as the view above suggests, equivalent or analogous to negative freedom. Those who are concerned with autonomy maintain that we ought to be concerned with more than the options available for people to choose between (i.e., noncoerciveness). We need to attend also to their capacities for self-direction. (I develop this view of the nature of autonomy and, by extension, this second line of critique of Thaler and Sunstein's version of the compatibility claim in much more detail in the next section).

For some, these points provide the basis for the *incompatibility view*. Emphasising the second part of the definition above – i.e., that nudges operate by making use of knowledge relating to the automaticity and context-dependency that affect human

behaviour and decision-making – critics contend that PCA is not compatible with the respect for individual autonomy that is due to citizens by their governments. This objection can take one of two different forms. The first (and most common) is the *objectionably manipulative critique*. As Mills writes:

> "Critics may suggest that choice architecture is necessarily heteronomous because it seeks to exploit heuristics and cognitive biases in our reasoning. Accordingly, choice architects pursue a programme of manipulation that undermines the independence of an autonomous agent's will by subverting the flaws in her decision-making competency to bring about particular ends." (Mills 2015: 498)

Choice architecture is, in the minds of these critics, a subtle and objectionable form of manipulation because it intentionally bypasses people's rational and deliberating faculties (i.e., their "System 2" thinking) in order to influence their choices by exploiting the various biases to which they are susceptible (Grüne-Yanoff 2012: 636; Hausman and Welch 2010: 128-129; Bovens 2009: 209; White 2013: 95; Wilkinson 2013: 347). The central thought is that it would be inconsistent to care about coercion but not about other kinds of treatment that share its worrisome properties. We care about coercion, when we care about autonomy, because coercion interferes with a person's independence. This being so, it would be "odd to care about coercion but not about manipulation", given that both kinds of treatment are objectionable on the grounds that they interfere with citizens' independence (Moles 2015: 654). Manipulation does this via a different means: it does not, like coercion, interfere with a person's options, but rather it "perverts the way that a person reaches decisions, forms preferences or adopts goals" (Raz 1986: 377-378). In so far as nudges operate by manipulating citizens' psychological processes without the consent of their target, they too offend against ethical independence. If, by its very nature, PCA falls foul of this objectionably manipulative critique, as many of these critics argue, it cannot be compatible with autonomy.

The second version of the incompatibility claim, which is related to yet distinct from the first, is based on what we might call the *objectionably paternalistic critique*, which charges PCA with being a disrespectful form of influence. Standardly, the key elements of paternalism are taken to be: "that an individual is prevented from, or manipulated into, performing an action or making a choice, on the basis that this will be in their own best interests, and usually the associated assumption that they are not the best judge of their interests" (Begon 2016: 356). The discussion above addresses the first element: though nudging is not coercive, i.e., does not prevent people from performing an

action, it is possible that it manipulates them into not doing so. For many, though, the distinctive wrong of paternalism is to be found in its *motive*, in particular, it is taken to involve a negative judgement about people's abilities to lead their own lives.[59] Jonathan Quong, for instance, defines paternalistic acts as those where (1) "An agent A attempts to improve the welfare […] of agent B with regard to a particular decision or situation that B faces", and (2) "A's act is motivated by a *negative judgement* about B's ability […] to make the right decision or manage the particular situation in a way that will effectively advance B's welfare" (2010: 80, emphasis in original; cited in Moles 2015: 651). On Thaler and Sunstein's view, nudging *is* motivated by an acknowledgement of the evidence that appears to show that people are not the best judges of their interests (or at least of the means by which to achieve these interests). It is often the case that this has been expressed in negative terms, referring to nudges as interventions designed to help us overcome the welfare losses produced by "predictable irrationalities", "blunders", "bad decisions", and by the "flawed" and "systematically wrong" judgements that result from "human frailty" (Thaler and Sunstein 2009: 19, 5, 79). As a result, some have argued that, in so far as they are motivated by such negative judgements about citizens' abilities to lead their own lives, nudges are objectionably paternalistic and so are incompatible with respect for autonomy.[60]

§4.1.2: Conditional Compatibility Views

Neither the compatibility view nor the incompatibility view, as they are presented above, is fully correct; rather, both capture elements of the truth, while simultaneously missing important insights. A third kind of view seeks to occupy the conceptual and normative space between these first two views in order to combine the insights, and avoid the problems, of each. The central thought is that different types of compatibility claim can be made. In the version above, it is claimed that nudges are always compatible with autonomy (or at least that they always adequately accommodate respect for

[59] For some examples of the motive-based characterisation of paternalism, see Shiffrin (2000), Quong (2010), and Tsai (2014). Begon (2016: §3) offers a good overview of motivational accounts.

[60] Building on fn. 58, it is worth noting at this point that there is a broad consensus in the literature that a "limited, carefully defined range" of paternalism is potentially justifiable (Husak 1980: 27), and that often included within this range is a form that "preserves and enhances for the individual his ability to rationally carry out his own decisions" (Dworkin 1971: 125). Arguably such policies would not be motivated by the types of negative judgements that anti-paternalists find disrespectful and, if so, they would not fall foul of the objectionably paternalistic critique. They would not only be compatible with autonomy, they would supports its development and exercise. This is one strand of the conditional compatibility view that I briefly sketch in the next subsection. This issue is discussed in full in §4.4.4, when defending the permissibility of autonomy-supporting nudge paternalism.

autonomy).[61] But we might also claim that PCA is only compatible with autonomy under certain conditions. As a general category, claims of this sort represent the *conditional compatibility view*. There are several ways in which this view might be specified. The aim of this chapter is to outline what I take to be the most attractive version of this selective view.

The general contours of the view that I'll be defending are as follows. It agrees with nudge advocates that certain types of PCA can be compatible with autonomy and, in fact, can act as a means of autonomy support. Nevertheless, it offers a much more discriminating view of this compatibility, which picks out only certain kinds of good as permissible for nudges to serve – goods that are specified by engaging fully with the autonomy-based concerns raised by critics. In doing so, it agrees with these critics that certain forms of choice architecture are impermissible on account of their effects on (i.e., their incompatibility with) citizens' autonomy. This marks the conditionality aspect of the view: nudges are morally permissible only in so far as they meet the conditions of operating in non-manipulative and autonomy-supporting ways or, in negative terms, in so far as they are neither manipulative nor objectionably paternalistic. The remainder of the chapter outlines and defends this view.

Its first task is to discuss the nature of autonomy in more detail, and to present an account of autonomy that adequately captures and explains both the internal and external threats to this ethical ideal. This is pursued in the next section. This account will provide the basis on which to assess the extent to which PCA is manipulative or not (§4.3) and, as a result, whether there are types of nudges that are able to support citizens' autonomy (§4.4).

§4.2: Autonomy, Automaticity, and Policy Interventions

The first important issue raised by this discussion is that there are different conceptions of autonomy at work in the debate over nudging, and that this is a source not only of conceptual confusion but, in the case of Thaler and Sunstein, of problematic normative argumentation too. The literature on autonomy is both vast and complex, comprised as it is of many competing accounts of how this concept ought properly to be understood (e.g., Arpaly 2002). At the broadest level, we might distinguish between two types of

[61] As Riley writes, Thaler and Sunstein "do not simply hold that nudging is in some circumstances morally permitted and practically called for. Rather, they favour and would have us foster [their] broad adoption […], to be deployed as a general purpose tool for good, across the institutional milieu of contemporary social life" (2017: 2).

conception – sometimes referred to as "choice autonomy" and "agent autonomy" (Lillehammer 2012: 196-204) – that parallel the distinction mentioned above between negative and positive freedom (Berlin 2002).[62] Our focus here is on agent autonomy – more commonly known in the literature as *personal autonomy* or *self-governance*.[63] This is the case for three reasons. The first, and most obvious, is that this is the kind of conception that was defended in the previous chapter as being a component of living well and which, as a result, plays a central role in our normative analysis of PCA. This reason in itself is sufficient justification, but two others add further weight. The second is that the majority of contemporary philosophical work has centred on analyses of the nature and normativity of this conception of autonomy. The third is that agent autonomy does not fall foul of the issues briefly outlined above that showed the choice autonomy view to be inadequate for the purposes of assessing PCA.

§4.2.1: An Account of Personal Autonomy

For our purposes it is not necessary to flesh out and defend a full theory of autonomy and, in any case, such a task would go beyond what is possible within the limitations of this thesis.[64] However, to make this concept – and this ethical ideal – more determinate, as is required for our normative analysis, we need to identify its constituent elements. One way of doing this, in the first instance, is to delineate three of its key intuitive features:

(1) *Capacity*: Discussions of personal autonomy take their lead from the idea that people have a capacity for self-directing activity, and that they exercise this to some degree.[65]

---

[62] Lillehammer defines these conceptions thus: "On my first interpretation [choice autonomy], 'being autonomous' means not being subject to coercive interference. […] On my second interpretation [agent autonomy], 'being autonomous' means being a self-governing agent" (2012: 196-197). If it were not for Thaler and Sunstein's adoption of the former view, and their referring to this as "autonomy", it may not have been necessary to draw attention to this distinction; but, as it is, it is important to do so for the sake of clarity.

[63] In what follows, I will use the term *autonomy* to refer to personal autonomy, and will use the terms "(personal) autonomy" and "self-governance" interchangeably.

[64] This task is taken up in numerous book-length contributions to the literature, including Dworkin (1988); Mele (1995); and Christman (2009b).

[65] There are some cases in which individuals do not have this capacity, such as when a person is in a coma or has a very severe case of intellectual disability. Many other cases are not so clear, such as in cases of "mental disorder" (Radoilska 2012). There are important questions about mental (in)capacity and its relation to the concept of personal autonomy. In this thesis, I am focused only on a central range of cases, namely, those concerning adult citizens assumed to be above a minimum threshold of autonomy. (*Capacity* also raises an interesting question about the possibility of autonomy in light of the empirical psychological evidence outlined in Chapter 2. I address this later in this section.)

(2) *Respect*: The possession of this capacity is thought to demand respect, such that there should be a presumption that an autonomous individual be allowed to exercise his or her autonomy and not be coerced or manipulated into serving another's projects.
(3) *Value*: Developing or applying this capacity further is considered to be a good thing.

Our first task in this section is to give an account of *Capacity*. Broadly speaking, personal autonomy refers to the capacity, usually considered to be special to humans, the possession and exercise of which enables an individual to be a self-governing agent. There are many different (though often "overlapping") accounts of what is required for self-governance. The issue at stake concerns how we should understand the autonomous agent's special relation to her own motives. Sarah Buss (2013: §2) distinguishes between three main types of account in the literature: (i) *coherentist* accounts maintain that an agent is self-governing if and only if "she is motivated to act as she does because this motivation coheres with some mental state that represents her point of view on the action" (e.g., Frankfurt 1971; Watson 1975; Bratman 2005); (ii) *reasons-responsiveness* accounts hold that an agent is self-governing if and only if "her motives, or the mental processes that produce them, are responsive to a sufficiently wide range of reasons for and against behaving as she does" (e.g., Fischer and Ravizza 1998); and (iii) *responsiveness-to-reasoning* accounts, the type of view that I favour and outline in more detail below, emphasise the importance of the reasoning process itself to self-governing agency (e.g., Christman 1991, 1993, 2017; Mele 1995).

According to the view I favour, then, the relationship between an individual and her motivational states is characterised, most broadly, as her ability to decide (or to have the appropriate control over, in some other way) which of them to follow (Arpaly 2002: 118-9). We might introduce some of the aspects of this view by referring to Berlin's characterisation of positive freedom, which derives from the capacity and the wish of an individual:

> "…to be a subject, not an object; to be moved by reasons, by conscious purposes, which are my own, not by causes that affect me, as it were, from the outside. I wish to be […] a doer – deciding, not being decided for, self-directed and not acted upon by external nature or by other men as if I were a thing, or an animal, or a slave, incapable of playing a human role, that is, of conceiving goals and policies of my own and realizing them." (Berlin 2002: 178)

The essential claim is that self-governing agency rests on the capacity of an individual to evaluate her motives on the basis of her own beliefs and desires, and to adjust these motives in response to these evaluations (Buss 2013). In Berlin's terms, it is the capacity to be moved by her own reasons and conscious purposes, so as to be able to conceive goals and policies of her own and to realise them in action. It is clear that, on this view, it is the autonomous agent's responsiveness to her own reasoning that matters centrally.

The repetition of "my own" in the phrases "moved by reasons, by conscious purposes, which are *my own*" and "conceiving of goals and policies of *my own*" highlights two important aspects of this view. The first is that, unlike reasons-responsiveness accounts, a person can exercise this capacity for self-governance even though her own reasons and purposes may include false beliefs about what she has reason to do. Personal autonomy, as an ethical ideal, "refers to a life of partial self-creation" (Wall 2012: 101; see also Dworkin 2011 and Raz 1986: 369), and one way of thinking about what this means is for an individual's attitudes, beliefs, values, goals, etc. to be "at least in part up to her", such that she is able to have a say in creating or inventing the kind of person it is worth her being (Helm 2001: 12). In conceiving certain "goals and policies", then, a person determines for herself many of the prudential reasons that apply to her. Raz calls this process, which marks a distinctively human kind of freedom, the "self-creation of value" (1986: 389). I take it that this ideal of self-authorship is, in one sense at least, what Berlin means when he makes reference to being moved by reasons "which are my own".

The second aspect is that responsiveness-to-reasoning views, unlike both coherentist and reasons-responsiveness accounts, are inherently sensitive to the intuition that certain external influences undermine our autonomy. In the case where a person's practical reasoning is directly manipulated by others ("being acted upon […] by other men"), this person would not govern herself by means of this reasoning and so she would have no power over the motives that this distorted reasoning produced (Buss 2013). This attests to the important idea that autonomy "is in part a social ideal" (Raz 1986: 378). It is not only the consequences of coercive and manipulative treatment that matter, i.e., the distortions that they cause in terms of achieving coherence with higher-order attitudes or being sufficiently responsive to what one (really) has reason to do. Violations of the independence condition matter also "because of the kind of treatment of others that they are" (ibid.). Treatments such as manipulation are fundamentally

"designed to restrict rather than enlarge conscious deliberation" (Barber 1971: 66), by short-circuiting people's own critical and practical reasoning processes.

There are, therefore, several reasons to favour responsiveness-to-reasoning accounts, and more specifically what we might call *autonomy-as-self-authorship*, over competing accounts. I will outline two now, and a third in the next section. First, at a general level, this conception is able to capture a fundamental element of the nature of personal autonomy which eludes coherentist and reasons-responsiveness accounts. These two accounts both conceive of self-governance as a kind of "constraint": to be autonomous on the former, one's motivations must be organised in a certain way; to be autonomous on the latter, one must act only on certain reasons (Colburn 2010: 13). By contrast, in the case of autonomy-as-self-authorship, there is an important sense in which the particular conditions that determine whether or not a person is autonomous are, at least in part, decided by the agent herself.

Second, a further advantage is that this conception offers an attractive combination of internalism and externalism (Mele 1995: 146-149), which remedies the issues caused by accounts that are either too internalist (i.e., coherentism) or too externalist (i.e., reasons-responsiveness) in character. Like coherentism, autonomy-as-self-authorship maintains that autonomous agency requires the ability to take a more reflective standpoint on one's motives, and that motives authorised from this standpoint are *internal* to the agent herself in a way that other motives are not. But autonomy-as-self-authorship provides a more plausible account than coherentism of this capacity for self-reflection. In the case of the latter, this capacity is comprised of the ability to hold higher-order attitudes; but this, in itself, fails to explain the authority of such attitudes in grounding self-governing agency, which, it is argued, comes from the practical reasoning that supports these higher-order attitudes. Identifying with a motive is important to autonomy in so far as "identification" is representative of implicit claims about which motives are supported by the agent's reasoning. On the other hand, like reasons-responsiveness theorists (and unlike strict coherentists), the reasoning-responsive view contains an important externalist element. Namely, it recognises that whether or not an individual's actual practical reasoning counts towards self-governing agency is dependent upon "which forces exert a nonrational influence on this reasoning" (Buss 2013).[66] An attractive feature of autonomy-as-self-authorship, though,

[66] It is therefore similar to (one reading of) Gerald Dworkin's account, which adds a "procedural independence" condition to his otherwise internalist view. He writes, "Second-order reflection cannot be the whole story of autonomy. For those reflections […] may be influenced by other persons or

is able to build in an *external* condition without requiring that autonomy involves an epistemically demanding relation between the agent's attitudes and external reality (i.e., an understanding of what she (really) has reason to do), as is the case in reasons-responsiveness accounts.

I take it that these reasons are sufficient, if not separately then together, to ground the adoption of autonomy-as-self-authorship as the conception of personal autonomy that will be used in the normative analysis of nudging. Hence, any reference to "autonomy" in the rest of the chapter can be assumed to refer to this conception, unless otherwise specified.

§4.2.2: Two Kinds of Threat to Autonomy

We are now in a better position to begin assessing the different types of compatibility claims outlined in §4.1. It is clear that, on this view of personal autonomy, nudging as a general purpose tool would be incompatible with autonomy in so far as it operates on its target in a manipulative fashion. It would fail the independence condition (and *Respect*, more generally) because, in disregarding an individual's reasoning capacity by seeking to pervert the way that she comes to make choices, it would bypass the central part of what makes the resultant choice *her own*. The next section is devoted to assessing fully this charge of manipulation and the extent to which it applies to PCA (§4.3). The focus here is to draw out in further detail why it is that Thaler and Sunstein's compatibility view is mistaken.

As Sunstein admits in an interview with *The New Statesman*, "There's a discussion of the ethical issues in *Nudge*, but we certainly didn't think about the issue of manipulation in any detail" (quoted in Lewis 2016).[67] We can see now that this is due to Thaler and Sunstein's "refusal to engage at any great length with autonomy as an intrinsically valuable consideration" (Mills 2015: 497, fn. 3). To unpack this, we need to understand the implications of their problematically thin conception of *Capacity* for both *Respect* and *Value*. Sunstein is aware that his view of autonomy (as freedom of choice) differs from other, more standard views in the literature. He draws a distinction between what he calls "thin" and "thick" conceptions, which differ primarily with

---

circumstances in such a fashion that we do not view those evaluations as being the person's own" (Dworkin 1988: 18).

[67] The only reference is in a brief discussion of possible objections at the end of the book, where they ask "What limits should be placed on […] public manipulation as such?" and answer by endorsing Rawls's "publicity principle" which amounts, essentially (and unsatisfactorily), to the view that a government "should not be secretive about what it is doing" (Thaler and Sunstein 2009: 244-245). More is said about this response in §4.3.3.

respect to the value they assign to autonomy. The thin conception, which he favours and wishes to defend, is "welfarist" (Sunstein 2014a: 126). This has two important aspects. The first is that the association between freedom of choice and well-being is taken to be *contingent*. As we mentioned above, Thaler and Sunstein present a fairly pessimistic view of human beings' capacity for practical reasoning and self-direction. They take the central lesson of the evidence from the cognitive and behavioural sciences to be that when we are free to choose, we do not always choose in ways that promote our well-being (Sunstein and Thaler 2003: 1198, fn. 143). Consequently, freedom of choice is valued only instrumentally, i.e., in so far as it promotes our well-being. The second aspect is that "freedom of choice is an *ingredient* of welfare, and when we decide what government ought to do [including whether it is permissible to nudge or not], we need to take account of the harmful effect, on welfare, of interfering with that freedom" (Sunstein 2014a: 124; emphasis in original). The key thought is that it simply is a fact that people generally have a preference to choose for themselves when it comes to how to lead their own lives.[68] Thus, on the thin conception, freedom of choice is valuable only in so far as it counts in a person's "welfare calculus" (Sunstein 2014a: 124-126).[69]

It is now possible to see what holding this thin conception of autonomy means for Thaler and Sunstein's normative analysis of manipulation (and, by extension, their account of the permissibility of nudging), and how this differs from the account of the wrongness of manipulation defended here. While manipulation generally tends to be ruled out by *Respect*, Sunstein acknowledges that the ethical analysis of manipulation is "much less straightforward" on his welfarist account (2016a: 85). For welfarists, what matters is how people's lives are going; and since manipulation might in fact promote people's well-being, understood in terms of informed preference satisfaction, they have no *in-principle* opposition to manipulation. This leaves it as an open question whether

[68] See, e.g., Sunstein's experimental work on public attitudes towards nudges (Sunstein 2016a, Ch. 6; 2016b).

[69] For Sunstein and others, such as Sarah Conly, "taking a welfare loss into account" means accounting for it within a cost-benefit analysis, because they adopt a broadly consequentialist approach to political morality. Sunstein, for instance, claims that "Social welfare […] is the master concept… If [interventions] are well justified on cost-benefit grounds, they should count as acceptable forms of paternalism" (2014a: 18-19). And, in a similar vein, Conly states that "Where choices should be left to the chooser, and where intervention is permissible, will be a function of what is best described as a cost-benefit analysis, rather than a decision a priori that certain personal decisions should be sacrosanct" (2013: 17). In principle, this analysis would permit keeping welfare-ptomoting manipulation hidden according to a "what people don't know won't harm them" rationale. In discussing this possibility, Sunstein mentions the risk of this type of strategy, namely, if hidden manipulation is disclosed then this may result in added welfare loss (relative to unhidden manipulation).

manipulation is wrong. Everything rests on the empirical consideration of whether a manipulative action in fact promotes or harms people's well-being thus understood (Sunstein 2016a: 99-100). As a result, such a position maintains that we can draw a distinction between harmful manipulation, on the one hand, and benign (or even beneficial) manipulation, on the other (Baron 2016: 283). This is summed up by Sunstein thus: "if manipulation really does increase welfare, then it would seem to be justified and even mandatory on ethical grounds" (2016a: 100).

By contrast, holding as I do the view that personal autonomy is intrinsically valuable, my view falls more into the category of what Sunstein calls the "thick" conception of autonomy. It is worth noting that, as Sunstein describes it, the thick conception differs slightly from the account that I have defended. To see this, we need to recognise that there are two senses in which *intrinsic value* might be understood, namely, either as valuable for its own sake or valuable in itself. As Steven Wall explains:

> "these two locutions are not equivalent: the latter refers to the source of value, implying that an intrinsically valuable thing gets its value from itself, not from something else; the former refers to the way in which a thing is valued, implying that an intrinsically valuable thing is valued for its own sake and not for the sake of something else." (Wall 1998: 145; for a similar point, see Korsgaard 1983)

According to Sunstein, the thick conception claims not that autonomy is "part of welfare but that it is an end in itself" (2014a: 127). Sunstein may or may not be aware of these two different locutions of intrinsic value; but, if he is, it would appear that he specifies the thick conception as conceiving of autonomy as valuable in itself.[70] By contrast, I have been careful both to describe personal autonomy as a *constituent* of a larger whole and to claim that its value derives from its contribution to this larger whole, namely, living well. Hence, in my view, autonomy does not have value in itself, but is valuable for its own sake, where the ultimate source of its value comes from its (significant) contribution to human well-being.

Nothing important hangs on Sunstein's view of the "thick" conception. I highlight this point to clarify the crux of the dispute between views like Sunstein's and views like the one I'm defending in this thesis. Manipulation is impermissible for Thaler and Sunstein if (and only if) it sets back a person's welfare interests, while, on my view, manipulation (at least that which is not consented to) is always harmful to a person's

[70] For more on the distinction between autonomy as right (or "demand") and autonomy as part of a person's good (or "benefit"), see Darwall (2006: 265) and Groll (2012: 706).

welfare interests, given the intrinsic value of autonomous agency. Like coercion, manipulative influences represent "an invasion of autonomy whose severity exceeds the importance of the distortion it causes" (Raz 1986: 378). Certain (i.e., manipulative) nudges are therefore morally impermissible even where they correctly identify (some of) the reasons that apply to their target and, as a result, in some respects would make their lives go better. This is because these nudges are set-backs to welfare interests that are also wrongs.

Sunstein makes reference to this type of view in a footnote in which he acknowledges that he is "bracketing here various questions about how welfarism is best understood. It is possible to have a conception of welfare that includes consideration of autonomy and dignity" (2016a: 98 fn. 47). Given that he does not offer any arguments against such a view, and given the arguments that I have presented in its favour, the remainder of the chapter assumes that autonomy is valuable for its own sake – i.e., not in a way that overrides well-being considerations, but in a way that it is *fundamentally connected* with these considerations about how well a person's life is going. Accordingly, to the extent that living well matters, any nudges that are shown to operate in an objectionably manipulative fashion are impermissible. In so far as this is shown to be the case, nudging is incompatible with autonomy because it undermines its externalist condition.

What the incompatibility view misses, however, is that there is a second relevant threat to autonomy, and that it might be possible for nudges to play some positive role in overcoming this threat. Clearly, the empirical evidence relating to automaticity and context-dependency poses a threat to the realisation of the ideal of autonomy-as-self-authorship: fast thinking and unreflective motivation may generally challenge a person's capacity to engage in practical reasoning or to realise the "goals and policies", to refer back to Berlin's words, conceived via such reasoning processes.[71] This fact is important and will be treated as such in my normative analysis.[72] But it is important to be clear about what the evidence has shown. Although it has highlighted the extent of these

[71] See Blumenthal-Barby (2016: 9-13) for an instructive discussion on the impact of biases on autonomy.

[72] Not all automatically activated behaviour needs to threaten autonomy. As Bargh and Chartrand highlight, part of the way in which we acquire skills is to gradually automatise certain "intentional, goal-directed processes" so that they become more efficient over time and can, in the end, operate "without conscious guidance" (1999: 463). This process is conditioned by the "frequent and consistent pairing of internal responses with external events" (ibid.: 464). There is some discussion of whether such automatically activated behaviours can count as autonomous agency or not (e.g., Fenton 2014; Felsen and Reiner 2011); and much rests in this discussion on one's conception of autonomy. On my view, which is developed below (§4.4), conscious involvement is not always necessary *at the point of behaviour*. On a temporally extended view of autonomous agency, one can be responsive (at time *t*) to reasoning that has occurred at a previous point in time (at time *t-1*).

potentially autonomy-threatening cognitive processes, it is certainly *not* the case that it has revealed that autonomy is impossible (see Felsen and Reiner 2011). As Jennings et al. recognise,

> "it is a mistake to assume that just because contextual factors shape all of us all the time, and we all use cognitive and emotional short-cuts and make mistakes, it follows that the traditional [liberal] goal of becoming more reflectively self-aware [and self-governing] has been shown to be moot or an illusion by behavioural research." (Jennings et al. 2016: 90).

As a result, I take the central – and in many ways, commonsensical – lesson of the evidence from the cognitive and behavioural sciences to be that the capacities required by autonomy-as-self-authorship do not necessarily come as the default setting for humans. In response, and in contrast to the perspective taken by Thaler and Sunstein, it is possible to recognise that by diagnosing the various cognitive and motivational constraints that challenge the realisation of autonomy, the cognitive and behavioural sciences in fact provide a way of helping us better to combat and control them. In other words, an awareness of how these constraints systematically affect our practical reasoning creates an opportunity to mitigate them (see, e.g., Gigerenzer 2014). Given autonomy's intrinsic value, we ought to respond to the evidence by using the policy levers available to us to support and promote it. And one of the ways that this might happen is through the use of PCA.

This brings us to the third advantage of the conception of autonomy-as-self-authorship, promised above. This conception is, as Colburn also recognises, "clearly an ideal with political relevance" (2010: 19), and so it enables those who hold it to use autonomy as the foundation for a liberal political philosophy committed not only to its protection, but also its promotion. Autonomy-as-self-authorship thus also captures the third of the intuitive features outlined above, namely, that developing or applying this capacity further is considered a good thing. The confluence of the discussions relating to the two threats to autonomy (the threat from manipulation and the threat from automaticity) therefore raises the interesting question of whether some kinds of PCA are able to mitigate these autonomy-undermining factors, while at the same time meeting the independence condition for autonomy. Answering this question is the focus of the next two sections.

§4.3: Assessing the Charge of Manipulation

The first part of this task is to assess to what extent, if at all, PCA operates on citizens in an objectionably manipulative manner. This section briefly clarifies some issues relating to the nature of manipulation (§4.3.1), ahead of engaging in the chapter's core critical assessment (§4.3.2 and §4.3.3), which generates the main arguments in favour of the conditional compatibility view and, in turn, provides the basis for developing my account of autonomy-supporting nudge paternalism in the following section (§4.4).

§4.3.1: Clarifying the Nature of Manipulation

Up until now we have discussed manipulation without specifying its nature in any real detail beyond the idea that it is independence-undermining on account of the perverting effects it wields on a person's practical reasoning and decision-making processes. This view intuitively captures the core of the concept, but it leaves a number of issues to be clarified, which are relevant to our inquiry into PCA.[73]

Three points are important. The first relates to a distinction between two types of manipulation, namely, *situational manipulation* and *psychological manipulation* (Faden and Beauchamp 1986: 362-368).[74] This parallels the two ways in which the term is used in common language: (i) changing, altering, or editing something so that it is perceived differently, or (ii) covertly controlling or exerting influence on a person's psychological processes (where the former tends to be non-moralised and the latter moralised).[75] PCA is, by its very nature, a form of situational manipulation, since nudges are designed to modify in some sense their target's perception of a given situation. For example, the use of social norms in official letters – such as the inclusion of a local comparison statement in letters from the UK's HMRC (e.g., "The great majority of people in [the taxpayer's local area] pay their tax on time") – is designed to affect the receiver's perception or

[73] There has been a surprising lack of sustained philosophical analysis on the topic (Noggle 1996: 43; Sunstein 2016a: 79). This is now being addressed, given that psychological discoveries and technological developments bringing questions about the ethics of influence to the fore (e.g., Coons and Weber 2014).

[74] This takes its lead from Faden and Beauchamp's discussion, but differs from it in two important ways. First, Faden and Beauchamp identify three distinct categories of manipulative influence, but I am here discounting the first of these – which they call "manipulation of options" – because nudging (by definition) leaves options unaltered. These authors include in this category the modification of options through the offering of rewards or by threatening punishment, giving the example of "the manager who gets an employee to transfer to a new position by increasing the person's salary" thereby offering her an attractive new offer (1986: 355). Plainly, this would not count as a nudge. Second, they draw the distinction between "manipulation of information" and "psychological manipulation"; and although the former shares some characteristics with what I'm calling situational manipulation, it differs from it in so far as it is principally concerned with deception, whereas situational manipulation may or may not use deceptive means.

[75] See the OED entry, available at: https://en.oxforddictionaries.com/definition/manipulate.

understanding of the situation. Similarly, reducing the size of plates in a canteen changes people's perceptions about how much they need or want to consume (Wansink and van Ittersum 2013). But altering how an individual perceives a situation is not necessarily the same thing as psychologically manipulating her. These categories are analytically distinct because, even though situational manipulation can sometimes count as a form of psychological manipulation, it need not pervert the way in which people come to their choices – and it is possible, at least in principle, that it could support people's practical reasoning in various ways. According to Faden and Beauchamp, psychological manipulation occurs when an individual is "influenced by causing changes in mental processes other than those involved in understanding" (1986: 366).[76] For our purposes, and more generally, we might think of psychological manipulation as those kinds of influences that operate, to use Noggle's metaphor, by deliberately "adjusting someone's internal or psychological levers" in certain ways (1996: 47), thereby affecting the levers that guide our deliberation and practical reasoning.

It is psychological manipulation that concerns us normatively, on account of its autonomy-undermining effects on a person's reasoning. As noted above, situational manipulation as a means of influence may or may not operate also as a form of psychological manipulation; and it is precisely this hybrid category that we are interested in. Although he does not acknowledge this important distinction, we can see from Waldron's critical statement about the nature of nudging that this category is the object of our concern. On his characterisation, governmental choice architects "should *manipulate the choice architecture* so that those who are less likely to perceive what is good for them can be *induced to choose* the options that [the choice architects] have decided are in their best interest" (Waldron 2015; emphasis added). Thus, we can characterise the general category of "objectionably manipulative" that concerns us as psychologically manipulative situational manipulation.

A second point of clarification concerns intentionality. One issue in the debate over the nature of manipulation is whether it is a necessary condition of manipulative influence that it is done intentionally (see, especially, Baron 2014). My main reaction to this is simply to observe that I do not need to argue either for or against this position for manipulation generally, because the question we are interested in – whether choice architecture is manipulative – concerns a set of policy interventions that are by their

[76] The primary concern of these authors is informed consent, which explains why they define this form of influence in terms of its effects on "understanding".

very nature intentional, namely, paternalistically-motivated nudges.[77] Nonetheless, this response does not capture some of the more subtle aspects of this issue. To see this, we can notice that there is a difference between the two types of intention: '*A* intends *B* to do *X*' and '*A* intends *B* to do *X* unless *B* wants to opt out' (Wilkinson 2013: 352), where, in this case, *A* is the government (or, more precisely, a choice architect), *B* is a citizen, and *X* is a pre-determined behavioural outcome. It could be argued that public nudges fall into the second category of intention and that, if so, the government does not intend to manipulate its citizens, when it does (which, on accounts that assume intentionality as a necessary condition, would discount them as cases of manipulation). This might be correct, in principle. But this would require that the government sincerely wanted, and also sufficiently allowed for, people to opt out. In turn, this requires allowing people to have a genuine choice, which would require that the government avoid methods that bypass the deliberative processes that are connected to decision-making. Hence, if we find that certain forms of nudging operate by bypassing these processes, then they would count as sufficiently intentional and so governments would not have available to them this type of defence against such interventions being described as a form of objectionable psychological manipulation.

A third point relates to consent and, in particular, the possibility of consensual manipulation. We have specified above that psychological manipulation is generally impermissible on account of its effects on autonomy. But I have been careful to specify this claim with an element of conditionality, because, if psychologically manipulative nudges could be consented to in the appropriate way, they would not fall into the category of *objectionably* manipulative influence.[78] This raises the further question of what sort of consent would be required for this to be the case. This issue will be addressed in

[77] This does, nonetheless, point us in the direction of an interesting separate question relating to foreseeability. Does it matter morally whether an individual is nudged intentionally (for paternalistic reasons) or that their being nudged (in a paternalistic way) is a foreseeable side-effect of a policy that is pursued for other reasons? Take, for example, the difference between a choice architect reducing the size of plates in school cafeterias in order to reduce calorie intake, versus her buying smaller plates because they are cheaper (and the money that is saved can be spent on some other educational need), but knowing that this would have the effect of reducing calorie intake. The permissibility of foreseeably, but not intentionally, influencing people's behaviour requires further reflection, and is an issue that has, to the best of my knowledge, so far been neglected in the literature. I thank Kimberley Brownlee for drawing my attention to this issue.

[78] Wilkinson highlights that there is a question about "whether we should think of consenting as making a manipulative act not wrongful or making the act not manipulative at all" (2013: 353). Across all potential forms of manipulation, it seems possible – if not probable – that consent would prevent manipulation in some cases, while in others it would still count as manipulation but would not be objectionably manipulative. I need not take a stand on this issue, for our purposes. I simply accept what I take to be the less controversial of the two positions, namely, that any nudges that fall into the category of "appropriately consented-to manipulation" are not objectionably manipulative.

more detail below; but it is possible to outline two central aspects of the answer at this stage, given the nature of autonomy and its value to a human life defended above. First, there is no inconsistency in holding that a person's autonomy entitles her knowingly to limit her own agency in certain ways and to some extent, and that this can be done for the sake of protecting or supporting her autonomy. Second, manipulative public nudges can be made consistent with autonomy only when the individual herself consents, and "not when other people do, in a democratic majority or not" (Wilkinson 2013: 354; see also Sunstein 2016a: 105-107).

§4.3.2: (When) Is Choice Architecture Manipulative?

We are now in a position to tackle the chapter's central question head on: When are public nudges objectionably manipulative? Thus far we have established that nudges are impermissible *in so far as* they operate via psychological manipulative mechanisms. Critics object that nudging as a general purpose policy lever falls foul of this problem. A common characterisation of how nudges operate in the critical literature is that they "deliberately circumvent people's rational and deliberating faculties, and instead seek to influence their choice through knowledge of the biases to which they are susceptible" (Grüne-Yanoff 2012: 636; emphasis added). The first part of the argument in favour of the conditional compatibility view rests on noticing an important ambiguity in this characterisation's use of the term *choice*. On the one hand, it is common to regard a "choice" as the particular behaviour, among the set of the behaviour-options available to her, that a person in fact undertakes, regardless of the process by which this came about. On the other hand, "choice" is commonly used, especially within philosophical discussion, to refer to the behavioural outcome that follows from the deliberative process of a person making a decision, which requires some active deliberation on the options. We can call these two senses *choice-as-behaviour* and *choice-as-decision*, respectively. This draws on the two-system approach to understanding human cognition outlined in Chapter 2: the former, but not the latter, would include behavioural outcomes that result from the automatic processes of System 1 as choices.

It is clear from the way in which this common characterisation is formulated that critics mean choice-as-behaviour when they refer to "choice" because they depict nudges as interventions that necessarily and deliberately "circumvent people's rational and deliberating faculties". The core of the *objectionably manipulative critique* is that nudges are designed to push people in ways that "bypass [their] reasoning" (Conly 2013: 30). There are good grounds for raising this worry. PCA is designed specifically to affect our

automatic modes of thinking (i.e., System 1 processes). Nonetheless, our normative assessment needs to ensure that it does not adopt an overly simplistic view of the empirical research. To this end, it is crucial to recognise the *interaction* between the two cognitive systems. Kahneman is clear that his research is focused on "the workings of System 1 and the mutual influences between it and System 2" (2011: 13). While System 1 can operate independently (which accounts for the evidence of automaticity), System 2 operates on premises and in a context provided by System 1 thinking. So, even though nudges do not work *directly* on System 2 processes, it is possible to design choice architecture with the aim of engaging and/or influencing our more deliberative processes.[79] Some accounts that share the main thrust of the common characterisation may be able to capture this subtler view. Hausman and Welch, for instance, describe nudging as "the use of flaws in human judgement and choice to influence behaviour" (2010: 124). On the distinction sketched above, the "flaws in human judgement" are biases that affect System 1 thinking, and we might specify the "flaws in […] choice" as those that affect System 2 processes. In both cases, behaviour can be influenced or manipulated; but there is a difference between doing this at the level of conscious deliberation and decision-making, and doing so at the level of automatic judgements.

This first distinction, then, allows us to distinguish between choice architecture designs that aim at influencing choice-as-behaviour, which we can call *System 1 nudges*, and those that aim at influencing choice-as-decision, *System 2 nudges*.[80] An example of the former is reducing calorie intake, and thereby seeking to fight one of the main causes of obesity and its associated health risks, by decreasing portion sizes, packaging, and crockery. This change in the choice environment has been shown to affect people's behaviour – i.e., their choice-as-behaviour about how much to consume at a particular sitting – but it does not do so by engaging any deliberative processes. Instead it brings about this behavioural change by enlisting or exploiting automatic cognitive processes. Consequently, when people reduce their calorie intake, it is as a result not of their responding to their own reasoning relating to consumption choices, but rather their responding behaviourally to the choice architect's reasoning about how much they

[79] There is an important distinction here, which will be drawn in more detail below, about the difference between seeking to engage deliberation and seeking to influence it. For now, the main point is that, in such cases, nudges could use insights from the cognitive and behavioural sciences to affect choice-as-*decision*, and that this is something that does not fit with the idea that they seek intentionally to circumvent or bypass people's rational or deliberative faculties.

[80] For a similar distinction between *type 1* and *type 2* nudges, see Hansen and Jespersen (2013). Sunstein also makes a distinction between System 1 and System 2 nudges in his experimental work (though, as is explained below, we understand it differently).

should be consuming. Another example would be the use of knowledge of the so-called middle-choice bias to direct people towards the option that the choice architect considers to be the best one. There is evidence that people tend to avoid the highest and lowest priced of a set of options; and a government could use this to design choice architecture that positions, e.g., the choice architect's favoured health insurance option between a higher and a lower option as a means of taking advantage of this bias. Like the calorie-reduction technique above, this intervention seeks underhandedly to direct people's practical reasoning. A third example of a System 1 nudge is the use of default rules. The principle upon which defaults work is that, since people often do not think enough about certain important issues such as saving for their pension or registering as an organ donor (or, if they do, sometimes fail to put this into action), it is good practice for policymakers to set a default option that will direct people's behaviour in the way that is considered to be welfare-promoting. This decides on behalf of citizens how they will behave if or when they do not deliberate themselves. We see this in the case of efforts to change pension schemes from opt-in to automatic enrolment. The latter default is designed to take advantage of people's inertia and status quo bias, meaning that outcomes are often not the result of any deliberative process by the target of the intervention.

By contrast, an example (or, rather, a general category) of System 2 nudges is "active choosing" policies. Instead of automatically enrolling citizens into being organ donors, it is possible to design PCA that prompts citizens into considering whether or not to register as a donor at all and, if so, which organs they are willing to donate.[81] This provides people with the opportunity to respond to their own reasoning on this issue, rather than offering a "one-size-fits-all" default. Another type of example is presenting nutritional information on food items using the "traffic light system". This provides the same information as before, but does so in an accessible and noticeable way. This make it possible that this information will influence "the attention and premises of – and hence the behaviour anchored in – reflective thinking (i.e., choices), via influencing the automatic system" (Hansen and Jespersen 2013: 14).

This distinction allows us to make an initial point about what a concern for autonomy implies for choice architecture policy. At a broad level, autonomy supports

[81] For a full discussion of the different types of nudging in the organ donor registration case, see MacKay and Robinson (2016). This is a nonpaternalistic nudge. I make reference to it here simply to provide an example of a System 2 nudge, since this general distinction (between System 1 and System 2 nudges) is relevant to both strands of the normative analysis.

nudges that improve people's opportunity to make conscious and informed decisions, and it is hostile to nudges that merely aim to improve people's behaviour by bypassing their deliberative capacities. This captures an important part of the truth; but in itself it is too simplistic, because it is possible, in principle at least, for System 1 nudges to take on an autonomy-supporting form and for System 2 nudges to take on an autonomy-threatening form. First, a System 1 nudge could support an individual's autonomy if she *consents* to it as a means of constraining her future behaviour in ways that ensure that it aligns with her considered judgements. This can support her autonomy by improving her responsiveness to her own reasoning. Second, although generally designed to support or increase a citizen's capacity to exercise her own agency (Sunstein 2016b: 129), there is a distinction to be drawn between two kinds of System 2 nudge – one of which is not supportive of autonomous agency. To see this, we might draw upon a distinction between directive and nondirective forms of influence.[82] The relevant difference is between interventions that are designed with the intention of steering a person's conscious deliberations in particular, content-dependent ways (we can call these *directive* System 2 nudges), and those that steer an individual towards deliberating on and choosing between options in a content-independent manner, that is, where no option is actively promoted (*nondirective* System 2 nudges). The latter category clearly provides an opportunity for, and indeed encourages, a person to be responsive to her own reasoning, since it does not take a stand on which putative reasons are sound.

By contrast, there are concerns attendant on the use of their directive counterpart in public policy. The deliberate framing of information in certain ways, for the sake of steering a person's conscious deliberations in a content-dependent direction, can be psychologically manipulative. Not all framed information provision is necessarily manipulative. An example of when it is involves the framing of cancer risks in relative, rather than absolute terms. A government might directively steer the deliberations of its (female) citizens about whether or not to have a mammogram by telling them that participating in screening reduces their risk of dying from breast cancer by 20%. The absolute figures tell a different story, however: screening reduces a person's chance of dying from 5 in 1000 women without screening to 4 in 1000 with screening (Gigerenzer 2015: 362). And, given that screening also comes with risks of false alarms and

[82] I borrow this distinction from the literature on the philosophy of education. There is is used there to distinguish between teaching methods and practices that have the intention of persuading students about the truth or falsity of particular beliefs and disbeliefs and those with the intention of encouraging critical abilities and understanding (Hand 2008).

unnecessary biopsies (such that screening may be all things considered more risky than refraining from screening), the 20% figure manipulates the input into a person's conscious deliberations and adversely affects her ability to discern the reasons that apply to her. Using behavioural insights to direct a person's deliberation in a way that intentionally biases it in favour of a particular pre-defined option selected by the choice architect is manipulative. Such interventions share similar characteristics with not-consented-to System 1 nudges (consider subliminal messaging as an extreme example). In both cases, the influence (i) is the intentional action of an external actor; (ii) outcome-focused, or directive, in nature; and (iii) acts on people in a way that means that outcomes are subject to having a (partially) covert explanation, because the true third-person explanation for a given choice, or at least a significant part of such an explanation, is hidden from the first-person perspective (Colburn 2011).

§4.3.3: Two Kinds of (Presumptively) Impermissible Nudges

Consequently, on my view, there are two main groups of PCA that are considered to be presumptively impermissible, where this means that there are weighty reasons to avoid these types of interventions.

Most obviously, there should be a presumption against the use of System 1 nudges in public policy, unless the appropriate consent has been given by the individual herself. This is because these interventions undermine a person's independence in a way that fails to improve her capacity to appreciate and act on reasons that she identifies as her own. To clarify how expansive (or not) this presumption is, we need to return to the complex issue of what counts as appropriate consent in the case of consensual manipulation. Consent justifies certain instances of manipulation by allowing that these interferences do not involve a violation of personal autonomy. From this perspective, we can see that non-explicit forms of consent are problematic. Hypothetical consent, for instance, involves the judgement of another agent, a choice architect in this case, about when it is reasonable to believe that an individual *would have* consented had the opportunity to consent been present. One way in which this might be done is through a related type of consent called "collective consent", which assumes that manipulative interventions might be democratically authorised in some way (Sunstein 2016a: 106). An example might come from citizens' juries, which are a type of public consultation that bring together a hundred or so members of the public for a day with the purpose of briefing them in detail about a particular issue and asking this sample of the population for their views on how they think the government should proceed. It might be assumed

from this process that, if a majority of the citizens' jury would consent to manipulative interventions, there is hypothetical consent (or "public permission") from a majority of the population at large (Halpern 2015a: 330-331). There is, of course, value to such democratic practices; but it would be incorrect to think that an act of manipulation, i.e., manipulative PCA, can be made compatible with personal autonomy in this way. In order for System 1 nudges to be made compatible with autonomy, they would need to be explicitly consented to by individuals. In other words, they would need to be personalisable.

The problem is that it is not clear how this could actually be achieved at the public policy level, at least without introducing other weighty problems. At the personal level, it is quite common for people to predict future irrationalities and motivational shortfalls, and to put irrationality- and temptation-defeating tools in place. Under certain conditions, this might be supported by PCA (more on this below, §4.4.2 on "auto-nudges"). Nonetheless, there is an important difference between these *personalisable* nudges and the *personalised* default rules that Sunstein claims are the "the wave of the future" (Sunstein 2013a: 10, 57). The latter are decided upon by choice architects, rather than being shaped by individuals. And this process of personalising default rules to different groups of people ("smart defaults") or potentially even to each individual citizen ("persistent defaults" and "adaptive defaults") raises serious concerns, which are separate from the problem of manipulation.[83] These more extreme, individualised techniques are becoming increasingly feasible due to technological developments that have made it possible, and indeed now common, not only to collect huge amounts of data but to "mine" this data to ascertain personality traits (Chittaranjan et al. 2013). They would work by tracking a person's past choices and creating default rules for her that are either based directly on those choices, or that involve a degree of extrapolation from those choices to other domains. If sufficient data is not available, it is still possible to use demographic and geographic data to create personalised defaults, whereby "[t]he general idea is that your default rules would track what would be best for 'people like you'" (Sunstein 2013a: 55). There are further autonomy-based problems with these proposals (e.g., the Millian idea that the creation of what we might call personalised echo chambers undermines learning and development processes that are associated with being exposed to different possible options), as well as weighty independent concerns relating to information acquisition

[83] For explanations of the three different kinds of personalised defaults, see Goldstein et al. 2008.

and privacy (especially in the case of the individualised defaults) and discrimination (especially in the case of the demographically-defined defaults). I take it that these points highlight that any such efforts to try to make default rules compatible with personal autonomy are not only likely to fail, but will also introduce additional reasons not to adopt System 1 nudges.

The second kind of presumptively impermissible nudge is the category of directive System 2 nudge that seeks deliberately to present information in a way that undermines the conditions for informed decision-making and responsiveness to one's own reasoning – as in the mammography example above. Like non-consented-to System 1 nudges, this category of influence objectionably manipulates people's practical reasoning processes. There are two main aspects to this objectionableness. The first is that these interventions take a stand on people's behalf about which specific putative reasons are sound or should be considered as weightiest in their deliberations. This is problematic within the prudential domain (or in the realm of what Halpern (2015a: 332) calls "lifestyle policy"). It means that they fall foul of the *objectionably paternalistic critique* due to their expression of a negative judgement about an individual's ability to make the "right decision" or to manage the particular situation in a way that will effectively advance her own well-being. The second, compounding aspect is that they express this stand in an underhand manner. Whereas traditional paternalistic interventions have taken on the form of mandates and other types of coercive policies, the two sub-categories of System 1 and System 2 nudges identified as presumptively impermissible fall foul of the *objectionably manipulative critique*, because they represent clear "risks to an agent's control over her own deliberation" (Hausman and Welch 2010: 136).[84]

[84] Can these nudges properly be said to threaten autonomy? One might argue that small denials of specific actions that are not attached to one's central goals and relationships may not affect one's autonomy at all (Raz 1986: 409-10). This may be true; in fact, it probably is. But psychologically manipulative nudges do not work by denying specific actions, i.e., by eliminating one option from the available option-set from which someone is able to choose. Rather, they adjust a person's internal, psychological levers in order to steer their behaviour in a particular pre-selected direction. Sometimes this will be in ways that do not relate to a person's central goals and relationships; many nudges, for instance, are directed towards influencing people's everyday behaviours. But many nudges do in fact affect important life decisions that might have considerable effects for people's central goals and relationships. For example, we have been (or will be) discussing nudge interventions that refer to cancer screening, health insurance schemes (particularly significant in the US context), pension plans, decisions about whether to have surgery, gambling addictions, whether or not one commits fraud, and so on. Manipulative nudges used in these domains cannot appeal to the idea that these are insignificant, everyday issues. But, to be clear, even in cases where the nudges affect more iterative, everyday decisions, such as food consumption choices, driving speeds, and so on, it would still affect people's autonomy if these were designed in psychologically manipulative ways – especially if these kinds of interventions are pervasive throughout society, as is increasingly becoming the case.

Some might respond by questioning why, especially in the case of health policy for example, we should insist on autonomy as our main guiding value. They might say that health is a case which shows the limits of autonomy as a value, because, simply put, it is more important to live longer than it is to have an informed and reflective choice about one's medical treatment. Isn't it better for people overall, for instance, that they are manipulated into being screened for breast cancer, than possibly dying from it due to missing out on the opportunity for early detection (because of, e.g., inertia)? My main type of response to this question is to challenge the way that it sets up the problem. It is true that autonomy is only one value among many, and that it is possible that it could be defeated in certain cases – and perhaps in cases involving health care – because its value does not override all other well-being considerations. Assessing this issue properly would require a comparative assessment of the effectiveness of manipulative and non-manipulative nudges (and other policy tools), and a value judgement on whether any additional effectiveness on the part of the former (if this is indeed shown to be the case) defeats the autonomy considerations. This judgement will rest on empirical information. But, given both that autonomy is a weighty value and that there will often be autonomy-supporting nudge alternatives that are likely not to be considerably less effective than their manipulative counterparts (if at all), we might assume that such cases will be quite rare and, hence, that public policy ought not as a general rule to sacrifice autonomy.

Another possible response, and one that has been offered by some in the literature, is that a transparency condition might alleviate the autonomy-threatening nature of these sorts of interventions. Thaler and Sunstein toy with this idea themselves as a response to the problem of manipulation (2009: 243-246). However, it is evident from both their brief discussion and the debate that has taken place since (e.g., Bovens 2009; Wilkinson 2013; Blumenthal-Barby 2013; Hansen and Jespersen 2013) that whichever way one specifies this transparency condition it is not sufficient to address the autonomy-based concerns we have been discussing. Blumenthal-Barby makes this point in the following way: "disclosure and transparency is not enough to preserve [autonomy] since manipulation […] occurs not just when a person is influenced without knowing it, but when they are influenced via exploitation of the non-rational elements of their psychological makeup" (2013: 191). So, even if we would expect a transparency condition to accompany any use of nudging, such a condition clearly is not sufficient to justify manipulation – a point that Sunstein now explicitly recognises (2016a: 104).

We can think of this subsection as the negative statement of the *conditional compatibility view* that I am defending. The next section clarifies the contours of the view more precisely by stating it in positive terms.

§4.4: Autonomy-Supporting Public Choice Architecture

Contra the incompatibility claim, our inquiry has found that PCA is not necessarily problematic from the point of view of autonomy. In fact, it is possible that it might, in certain cases, be able to work as a form of *autonomy support*.[85] There are two categories of PCA that may be able to play this autonomy-supporting role: the main one is nondirective System 2 nudges and the second is consented-to System 1 nudges. This section outlines both types of permissible choice architecture (in §4.4.2 and §4.4.3, respectively) and addresses the concern relating to the paternalistic nature of autonomy-supporting choice architecture (§4.4.4). Before this, though, I wish to spell out in further detail the ethical rationale for interventions of this kind.

§4.4.1: Justifying Autonomy Support

Recall that we delineated two relevant threats to autonomy: the external threat from manipulation and the internal threat from cognitive bias that affect our reasoning processes. If there were a means of reducing the internal threat, we can assume that the political community would have normative reasons to support these kinds of policies. The hasty shift towards a wholesale acceptance (or rejection) of "liberal paternalist" policies has meant that policymakers have failed to recognise the opportunity presented by the empirical research, namely, the possibility of using these cognitive and behavioural insights to enhance the conditions necessary for autonomous decision-making (Mitchell 2005: 1254-1257) and, as a result, for (one of) the conditions required for living well. As Rowson puts it:

> "The purpose of engaging with neural and behavioural sciences is to move away from a naïve perspective of freedom as unfettered individualism, towards an autonomy that is

[85] Nagel and Reiner (2013) also use this term, but not necessarily in relation to nudging. For them, autonomy support refers to "a form of decisional enhancement – manipulations akin to traditional cognitive enhancement with the objective of improving the outcome of decisions – but one in which the improvement is more unambiguously aligned with the wants, needs, and desires of the decision makers themselves" (2013: 36-37). The way I use the term is similar, but is focused primarily on supporting the procedural aspects of autonomy, in particular, practical reasoning. For other views of how PCA might support autonomy, see Mills (2015) and Binder and Lades (2015).

> grounded in informed self-awareness, including a deeper recognition of the social and biological conditions that define our actions." (Rowson 2011: 19)

Such a view needs to recognise that "[t]here is more one can do to help another person have an autonomous life than stand off and refrain from coercing or manipulating him" (Raz 1986: 407). Of course, these negative duties of non-interference are important; this is why the first part of the normative analysis has determined which forms of choice architecture are manipulative, so as to rule these nudges out as (presumptively) impermissible. But such an account tells us only what we ought *not* to do if we wish to respect people's autonomy. As we recognised above, autonomy-as-self-authorship also provides a basis from which to justify the creation of the conditions for autonomous agency. Such agency is dependent on having certain cognitive and motivational capacities, and being able to exercise them. Although we do have *Capacity*, the empirical evidence (and the inclusion of *Value*) highlights that its development and exercise is not simply a given for humans. Riley recognises this when he states that:

> "We neither arrive in the world with this capacity, nor do we simply receive it as a constituent or effect of sheer physiological maturity. Its development in one of us is, rather, a contingent culturally-mediated social and individual accomplishment. Further, it also appears that reasoning capacities of this sort need the right kind of setting, stimulus and ongoing support in order both to develop and to stay in good working order." (Riley 2017: 8)

This idea that autonomy is a (partly) social achievement is an important one. Perhaps the role for PCA in the development of personal autonomy is likely to be fairly minimal, relative to other kinds of interventions. But the point that I wish to defend is that PCA *can* have a role to play in providing the "right kind of setting, stimulus and ongoing support" for citizens' exercise of their capacities for self-direction. It can do so by using the knowledge of cognitive bias and motivational shortfalls to design PCA that negates these internal threats to autonomy.

This is all the more important within a context that has, in recent years, made this kind of supportive intervention more urgently needed. The same psychological evidence is being used by others within the political community to optimise their ability to capture our attention and to influence our beliefs and behaviour – the most problematic group of which Halpern refers to as "behavioural predators" (2015a: 312-313). In addition, the general proliferation of information and media has "made it

harder for people to be intentional or reflective about their goals and priorities in life" – a phenomenon that has been referred to as a "crisis of distraction" (Williams 2013). Given that these conditions threaten the development and exercise of autonomy, adequately satisfying *Respect* requires first recognising that autonomy has "social and political conditions", and then providing ways of designing (and regulating) these social and political conditions in ways that support it.

This is importantly distinct from traditional autonomy-based liberal views. Most liberal thinkers tend either to assume that (adult) citizens are sufficiently able to reason and to translate this reasoning into action in the way required for autonomous agency, or they claim that the government should treat them *as if* they are able to do these things, regardless of whether this reflects the reality of the situation or not. However, the cognitive science literature implies that these hands-off versions of autonomy-based liberalism need either to be updated so that they treat seriously the threat of automaticity to the realisation of personal autonomy, or to be replaced by a conception that is able to do this. Some have indeed taken the second, more extreme route. Susan Hurley, for instance, argues in favour of moving towards "an ecological conception of liberalism", contending that:

> "we should not try to shore up a traditional liberalism […] against the winds of empirically based change in our understanding of human minds. This nostalgic strategy does not provide much guidance to hopes for a liberal future, as the cognitive sciences continue to advance." (Hurley 2011: 215)

By contrast, this thesis takes the less (theoretically) ambitious first route. The remainder of this section outlines an account of autonomy-supporting PCA that treats seriously the threat of automaticity, arguing that, given the value of autonomy to a human life, the political community has normative reasons to support these kinds of nudge policies. That is, it argues both that state autonomy-supporting nudging does not wrong its target, and that there are *pro tanto* reasons for the state to engage in it. There is a separate question about whether or not there is a *moral requirement* that the state engage in autonomy-supporting nudging.[86] Such a requirement is possible in principle, on my view, but, in line with the thesis's principal aim, I focus on providing an account of

[86] Some have argued that not providing these enabling conditions would, in some cases at least, be unjust. According to Riley, it would count as a particular type of epistemic justice, which he labels "reflective incapacitational injustice" (2017: 9). I do not make this type of stronger claim.

permissible PCA, since the assessment of moral requirement is partially dependent on further facts.

§4.4.2 "Nudge-Autonomy"

The first, and the main, component of the conditional compatibility view is those System 2 nudges that fall into the category of what I will call the *nudge-autonomy* mechanism. We can imagine a range of nudges whose explicit purpose is to promote self-conscious deliberation. The aim of such interventions is to enable people to make their own (sometimes more informed) decisions by activating the cognitive resources of System 2 thinking. The impetus for this type of intervention is captured by Waldron's (2015) sentiment: "I wish […] that I could be made a better chooser rather than having someone on high take advantage (even for my own benefit) of my current thoughtlessness and my shabby intuitions". There are two distinct but related ways in which PCA might support this process.

The first is via the use of *active choosing* policy designs, which nudge an individual towards deliberating on and choosing between options in a way that does not actively promote any particular option. It is possible to set up "choice prompts" (Mills 2015: 502) in policy areas where it is thought that people are likely to suffer from cognitive biases or motivational shortfalls, or where other influences may be exerting (potentially) unwanted pressure on their choices. Instead of using choice architecture to control people's automatic responses, therefore, active choosing employs it for the purpose of enabling citizens to regain control over their own practical reasoning processes and its resultant behaviour. Instead of automatically enrolling people into a health insurance plan, for instance, a public university might ask people when they start their contract to indicate whether they want health insurance and, if so, which plan they want – the default one or a plan that is more specific to their own needs and circumstances.

The second is via *salient information provision*, which works by presenting relevant fact-based information in a salient form with the aim of triggering deliberation and making a person's reasoning and resultant decisions more informed. In other words, information can be provided so that it is noticed and absorbed by System 1 in a way that directly engages or activates System 2 thinking. In such cases, our System 1 "calls on System 2 to support more detailed and specific processing that may solve the problem of the moment" (Kahneman 2011: 24), essentially asking it what should be done in light of this information. Consider again the example of the traffic light system for displaying nutritional information on food items. This works by presenting health

information in a format that makes it both easier to notice and evaluate. As a result, it can both improve people's comprehension and increase decision-making opportunities by activating people's deliberation about their food choices.[87] By engaging people's conscious processes, these System 2 nudges can help people better to recognise and consider the reasons that they believe apply to them and to translate these reasons into suitable behaviour. Contrary to a common criticism, such processes need not be cognitively burdensome.

The "Ambient Orb" also works in a similar way. This is a technology that utilises the ability of the brain to perceive information without any apparent cognitive load. It works by providing feedback on energy consumption by glowing red when lots of energy is being used and green when energy use is modest, where this range is set by the user herself (Thaler and Sunstein 2009: 206). The Orb creatively makes users aware of information that is otherwise invisible to them. This does not bring about any automatic behavioural changes, but rather provides System 1 with information that causes it to activate System 2 (in the case of it glowing red, at least). The user can then 'check' whether this data is problematic or whether there are good reasons behind it. In the case where the user deems that energy use is unnecessarily high, she is able to take measures to reduce consumption, perhaps by unplugging devices that are fully charged, turning off lights in rooms that are empty, switching off computers or other devices that are not currently, and so on. This has been shown to be effective at reducing energy consumption by 40% in peak periods.

§4.4.3 "Auto-Nudge"

The second component of the conditional compatibility view is the category of appropriately consented-to System 1 nudges, which I call the *auto-nudge* mechanism (shorthand for "autonomous-nudge"). Auto-nudges reverse the sequence outlined above: rather than using System 1 processes to engage System 2 thinking, the auto-

[87] Within obesogenic environments, there are good reasons to make nutritional information salient as a means of counteracting the fact that unhealthy choices are easier to make than healthy ones in these contexts. In 2009, "almost a quarter of adults in England were classified as obese and three in ten children aged between two and 15 were classified as either obese or overweight". As one expert (Professor Baldwin) puts it, "the explanation for [this rise in obesity] is plainly not to be found in a collapse of personal responsibility" (House of Lords 2011: 52-53). All of the witnesses asked by the House of Lords Select Committee on Behaviour Change agreed that "the evidence demonstrated that those labels which included traffic light colours were better understood by consumers than those without" (ibid.: 55). And this increased comprehension has been shown to affect consumer behaviour: two large supermarket chains, Asda and Sainsbury's, provided the Committee with evidence that the introduction of traffic light labelling had led to a decrease in sales of those products with red (denoting "unhealthy") on the label.

nudge mechanism uses System 2 thinking to predict future irrationalities and motivational shortfalls in order to put System 1-based nudges in place to ensure that behaviour is responsive to one's reasoning about a particular issue. The central idea is that an individual, by cultivating an "informed self-awareness, including a deeper recognition of the social and biological conditions that define our actions" (to recall Rowson's words), might be able to regain control over or reconfigure her automatic behaviour, thereby effectively training her automatic responses to come more in to line with her own reflective reasoning about how she wants to live. This process has been referred to as "proactive habituation" (Rowson 2011: 26). We do this in our personal lives, and now increasingly so with the help of technology; but it is also possible for PCA to provide ways of supporting this process.[88]

There are two distinct but related (and potentially overlapping) ways in which the auto-nudge mechanism might work in practice. The first, which we have touched on briefly already, is via *personalisable default rules*. Such tools are only now starting to be developed, but BI Ventures – a part of the BIT – has recently launched a service that fits into this category. "Promptable" is a behaviourally-informed texting platform that has been designed with the aim of boosting attendance and grades at further education colleges (within the UK educational context in particular). One function of Promptable is that it allows the students themselves to decide whether or not they want to nominate two 'Study Supporters', and who they would like these people to be. If they do so, these elected supporters – be they a parent, older sibling, other relative, a friend, and so on – receive personalised text messages from the college at regular intervals, which allows them to initiate conversations with the student about their studies. The additional social support triggered by these text messages has been shown to have a significant positive effect on both attendance and exam performance (Groot et al. 2017).[89] Importantly, it is a type of PCA that the person themselves can opt into and can personalise.

The second is *institutional commitment devices*. Commitment devices provide a means by which an individual can lock herself into a course of action which (she foresees or fears) she might not otherwise choose, but which aligns with her reasoning

[88] Institutional support may be helpful because it is often difficult to sustain this re-habituation process ourselves. Recent research has shown that there is a neurological basis for this. It is difficult for people to sustain enthusiasm for behaviour changes they initiate, such as a new diet or exercise regime, because the rapid-firing of the two groups of neurons (in the 'ventral tegmental' and 'substantia nigra pars compacta' areas of the brain) that are critical for reinforcing the formation of new kinds of behaviour on account of the dopamine they release, gradually fall back to baseline after a certain number of iterations. So although we initially get "dopamine rewards that reinforce the forming of a novel habit", after a short time this reward wanes, leaving us reliant on our "scarce and depletable" willpower (Rowson 2011: 26).

[89] See Appendix F for a list of example text messages to supporters (Groot et al. 2017: 36).

about how she would like to act in this given circumstance (Offer 2006). One example of this type of PCA already in use as an institutional commitment device is that of voluntary gambling self-exclusion. In the UK, this programme is called "Opt-in to Self-Exclude". This offers a simple-to-use multi-operator self-exclusion scheme for those "who think they have a problem with gambling and want help to stop".[90] It provides a means by which such individuals can take back control of their lives through the re-habituation of their automatic cognitive processes. It lasts a minimum of six months, which ensures sufficient time for this reprogramming process to occur. It is possible to opt-in to this scheme online, which reduces the various costs attached to doing so, and it can be personalised depending on the type of gambling activities one is wanting to control.[91] This type of policy offers a promising template from which to formulate other kinds of pre-commitment-based auto-nudges.

Auto-nudges, in one way at least, seem at odds with the responsiveness-to-reasoning conception of autonomy, given that this conception emphasises the importance of the reasoning process to autonomous agency. This conception does not, however, require that an individual is continually choosing in a reflective way. As long as an individual is consensually nudged into acting in a way that conforms with the outcome of an earlier reasoning process, it is consistent with their autonomy to be nudged merely to behave (a System 1 nudge) rather than being nudged into deliberative choosing (the nudge-autonomy System 2 nudges). To clarify and to relate back to the earlier point on transparency, mere knowledge of the fact that this is occurring is not sufficient. In the cases that I am describing here, an individual has herself acted in some way – by actively personalising a default rule or establishing some other kind of pre-commitment – so as to steer or constrain her own (future) behaviour. As Hurley puts it, informed or self-aware agents "can accept and use lower level rational influences as ways of acting on their own reasons, deliberately delegating control to handy automatic processes in ways that further their goals" (2011: 214). This recognises the fact that being responsive to reasoning requires not only exercising the capacity to reason, but also exercising "self-control to hold to his deliberate decision" (Mill 2005: 71). This form of "motivational scaffolding", if you will, works in the same way as "Ulysses

[90] This quotation is taken from the programme's website, http://optintoselfexclude.info.

[91] For example, one is able to find on one webpage the opt-in details for self-excluding from arcades, betting shops, bingo venues, casinos, and online (remote) gambling. (One problem is that if an individual has a general gambling addiction, there would be quite a lot of administration involved in joining all of these separately-run self-exclusion schemes.)

contracts", which are widely recognised as being compatible with autonomy (for discussion of this, see Elster 2000 and Radoilska 2012).

§4.4.4: Autonomy-Supporting Nudge Paternalism

The central aim of both nudge-autonomy and auto-nudge is to correct or eliminate some kind of cognitive bias or motivational shortfall so as to enable people to act in ways that are more responsive to their own reasoning, and therefore more autonomous, than they would have otherwise been. In contrast to the two impermissible forms of nudging, it is the agent, and not the choice architect, who is ultimately guiding her behaviour in both cases. Consequently, these forms of PCA are compatible with autonomy and secure the main thrust of the argument for the conditional compatibility view.

Recall, nevertheless, that there were two kinds of incompatibility view – one based on the *objectionably manipulative critique* and another on the *objectionably paternalistic critique*. In order to complete the argument in favour of my conditional compatibility view, it is necessary to demonstrate that the account of autonomy-supporting choice architecture that I have set out does not fall foul of either kind of critique. Evidently, the main focus here will be on the second critique, from objectionable paternalism, given both: (i) that we have assessed the charge of manipulation in some detail and did not find anything there to concern us with respect to what we are now calling the nudge-autonomy and auto-nudge mechanisms of influence, and (ii) that I am referring to my account as a form of nudge paternalism, which draws the question about whether or not it is paternalistic in an objectionable sense. It will come as no surprise that I will be arguing that autonomy-supporting nudge paternalism is neither objectively manipulative nor objectively paternalistic. Nonetheless, arguing against the potential criticisms that could be levelled at my account helps to clarify some of its more subtle details and to give the fullest statement of the view possible.

Let's address the concern from manipulation first. Situational manipulation in itself is not objectionable manipulation; it becomes objectionable if it includes elements of psychological manipulation (which have not been actively consented to). Some might worry that the nudge-autonomy mechanism (but not the auto-nudge mechanism) crosses this boundary by attempting to intervene with the way in which people arrive at choices. More explicitly, the concern is that it essentially tricks people into thinking harder about their behaviour. The thought here is that these types of nudges deliberately act on System 1 thinking in a way that manipulatively steers people's practical reasoning

processes towards reflecting on a given issue. After all, these interventions have been designed to be processed by filters in certain ways and not others (Levy 2017). There are two kinds of responses to this concern. The first is to highlight that, even if we accept that this is a kind of psychological manipulation, albeit in an attenuated form, it does not offend against, but rather is *supportive* of, autonomy – which is the value that provides the normative foundations for the concern about psychological manipulation in the first place. The second response, however, shows that we do not need to accept this charge: once one has an empirically sensitive understanding of the reasoning processes that autonomy-as-self-authorship seeks to protect and promote, it becomes clear that influencing the attention and premises of reflective thinking via influencing our automatic processes does not count as an instance of bypassing the deliberative capacities of agents in any way. As Levy highlights, these "filters" are not an obstacle to reasoning, but are "partially *constitutive* of reasoning in normal functioning adults", which means that "designing arguments to appeal to them is not bypassing reasoning, but appealing directly to it" (2017: 4).

This deals with the objectionably manipulative critique. But, in doing so, it may have clarified that the real problem is not to do with how the nudge-autonomy mechanism operates, but rather with the attitude towards people's agency that it represents. We have already noted that autonomy is compatible with certain forms of paternalism, and we can assume that autonomy-supporting nudge paternalism fits into this category (see fn. 58 and 61 above). What those who are concerned with paternalism tend to worry about is that paternalistic interventions, even when they are autonomy-enhancing, issue from a negative judgement on the part of the paternalist about the target's ability to exercise her own agency (e.g., Quong 2010; Shiffrin 2000: 214). Interventions based on such judgements, then, necessarily involve a disrespectful attitude and set up a relationship of superior (i.e., the paternalist, and in this case, the choice architect) to inferior (i.e., the target of the intervention). It is this distrust of an individual's agency combined with the assumption that their own judgement is superior that makes the paternalist's actions wrong. Do the nudges included in the account of autonomy-supporting nudge paternalism express this motivational wrong?

Some might worry that they do, arguing that the main motivation for this version of nudge paternalism is that, in some situations, people make poor choices and that, with the help of choice architects guiding them in various ways, they will be able to make better (viz. more autonomous) decisions. It is worth noting that there is an

ambiguity here between two versions of this challenge, in particular, between a *comparative* and a *noncomparative* reading (Moles 2015: 652). On the former, nudge paternalism would be considered a comparative wrong because it treats some people as having less ability to exercise their autonomous agency, due to their cognitive biases and so on, than someone else. This version of the objection can be addressed by noticing that there need be nothing in the choice architect's motivation that involves the claim that she is better than the intervention's target population in any respect. As Moles highlights, her alleged superiority "is fully explained by the (cognitive) situation in which she finds herself" (2015: 653). Her actions are motivated by the recognition that the targets of the intervention are "currently in a context or state where people (including [the choice architect herself]) happen to make frequent mistakes in instrumental reasoning" (Coons and Weber 2013: 13).

The noncomparative reading, in contrast, views nudge paternalism as wrong because it is motivated by the claim that a person simply lacks the relevant capacities of autonomous agency, without any further claim about other people's superior abilities in this regard (e.g., Carter 2011). This version is less forceful than its comparative counterpart, it seems to me. It can be addressed by noticing that *Capacity* does not specify the extent to which citizens are able to exercise the capacities required for autonomous agency – only that they can and sometimes do. It does not, for instance, specify that people are always, in every occasion, able to exercise the capacities of reasoning and self-control. If we accept a less demanding version, which takes into account the empirical evidence about the kinds of particular contexts in which people in general are likely to be subject to, e.g., a cognitive bias that makes it more difficult for them to be responsive to their own reasoning, then it seems that any negative judgement that may be made by the paternalist would not be a judgement about any particular agent but would be located, rather, "at the level of human cognitive architecture" (Moles 2015: 653-654).[92]

Accordingly, autonomy-supporting PCA is paternalistic, but not objectionably so. It does not express disrespect for people's ability to run their own lives; rather, it takes account of *human* cognitive limitations in a way that respects the value that comes with further developing and exercising the capacities relating to autonomy-as-self-authorship. As Begon highlights, in the wake of recent developments and trends in the literature on paternalism (including, but not limited to, arguments justifying nudge

[92] Arneson makes the same point: "The negative judgement involved in paternalistic action may be due to deficiencies acknowledged to be common to all of us" (2016: 609).

paternalism), there is no longer any straightforward debate to be had between the traditional rivalling factions of "paternalists" and "anti-paternalists" (2016: 355). The account I have developed is an example of this more complex conceptual landscape.[93]

The essence of paternalism is that "(1) it aims to close an option or make a choice for a person and (2) it is intended for the person's own good" (Clarke 2006: 117). Clearly, autonomy-supporting nudges are intended for the person's own good. They are grounded in, and justified by, the view that part of living well is to live one's life in an active sense, and that support for autonomous agency is valuable because it secures this good, as well as valuable in the instrumental sense that people are more likely to decide and act in ways that align with other elements of their well-being. So the issue that needs addressing concerns (1); let's see how the two mechanisms fare in this respect. Auto-nudges do close options, but do so as a form of voluntarily consented-to self-paternalism (or what Husak (1980: 43) calls "*paternalism towards oneself*").[94] And nudge-autonomy does, in a minimal sense at least, make a choice for an individual. The important point here is that the choice about how to act in a given situation is not made on behalf of this individual. Nudge-autonomy interventions do not involve, as Gerald Dworkin's definition of paternalism specifies, "a usurpation of decision-making", even though they might count as, in one sense, "interfering with the way in which they arrive at their decisions" (1988: 123). The choice that is made for an individual is the one that it would be better to engage in some sort of conscious deliberation about how she might want to act in this given situation. The interference, therefore, comes in the form of nudging the person towards active decision-making, rather than usurping her decision-making.[95] Within the standard conceptual framework, this makes my account a form of soft paternalism (Feinberg 1989).

---

[93] It is possible that this traditional paternalist framework is not the best way to describe the dynamics at work here. There is conceptual work being carried out that seeks to think through the implications of adopting a more socio-ecological approach to the conditions of personal autonomy, given that the classic paternalist/anti-paternalist framework seems unable to capture some of the more subtle complexities in these discussions. One suggestion has been to introduce the concept of *maternalism* (Specker Sullivan and Niker 2017). I do not pursue this any further here, but raise it to note that conceptual work is being done which may, in time, begin better to capture these more nuanced dynamics.

[94] Auto-nudges aim to do precisely the opposite of "preventing people from doing what they have decided", which is the relevant aspect of Dworkin's definition of paternalism (1988: 123; more on this below).

[95] Active choosing policies that require decision-making (rather than simply prompting it) are the most paternalistic in this sense, because they also rule out the option of "choosing not to choose" (Sunstein 2015). Sunstein calls this "choice-requiring paternalism" (Sunstein 2017: 12). Some might think that this stronger version of the nudge-autonomy mechanism has a hint of the "forced-to-be-free" paradox about it. This may give us a reason to favour weaker, prompted choice versions of active choosing. But, in certain circumstances, prompted choice designs may not be possible or feasible, and required choice designs would be desirable. In such cases, it is not clear that autonomy is undermined at all.

Someone might accept the foregoing analysis, agreeing that there are non-manipulative forms of choice architecture, that these might be used to support personal autonomy in certain circumstances, and that this would be a good thing; but they might, nonetheless, hold the view that it would be wrong for the state to play any active role in creating and maintaining the social conditions that help citizens better to exercise autonomous agency. This is the challenge from *anti-perfectionism*: it charges my view with stepping over the line of what it is appropriate for the state to do by acting on reasons that have to do with promoting particular activities or ways of life on grounds relating to their intrinsic value (Quong 2011: 27). It is true, and not a regrettable feature in my view, that it does this with respect to its aim to support personal autonomy.

It is important to recognise, nonetheless, that my account represents a kind of *procedural* liberal perfectionism, if you will – a view that is elsewhere described by anti-perfectionists as "restricted neutrality" (Clayton 1993: 105).[96] The interventions that it recommends do not rest on the idea that it is the government's responsibility to steer people in any particular content-dependent way. It recommends that interventions (and here the focus is on nudge-autonomy interventions) should be designed as forms of content-independent influence, which refrain from taking a stand on which putative reasons are sound. By this I mean that the government should not be motivated to establish something like the traffic light system to steer people towards making *particular* reflective and informed choices; rather, the autonomy-based reason in favour of this nudge is that it gets individuals to make more reflective and informed choices that are more likely to align with their own preferences and goals, regardless of whether it improves their chances of choosing healthy food.[97] It maintains that the ability and opportunity to choose for oneself "is a good that is independent of the wisdom of what is chosen" (Dworkin 1971: 117), and it is this good that nudge-autonomy seeks to promote. The identification of autonomy with responsiveness to one's own reasoning is therefore compatible with the idea that different ways of life are rational and valuable for different people, so that, as Swift points out, "a state helps its members towards

[96] On Clayton's view, restricted neutrality prohibits the state from promoting, on the basis of certain reasons, ideals of living well that involve a more determinate content than that they are chosen and pursued autonomously (1993, 2006). On this kind of view, it would be possible to understand my account as anti-perfectionist, because it seeks only to promote what Colburn terms the "second-order" value of autonomy (2010: 50).

[97] Of course, these may overlap (and it may even be likely that they will do so).

freedom […] by doing what it can to help them to live in ways that are rational for them, as the individuals they are" (2006: 84).[98]

§4.5: Conclusion

This chapter has provided an answer to the question: If we care about autonomy, is nudging for the sake of improving the target's well-being ever permissible? It has argued that nudges motivated by this concern are permissible, but only when they meet certain autonomy-based conditions: first, that they activate deliberative choosing without steering it in particular, content-dependent ways (the *nudge-autonomy* mechanism); and, second, if they work at the non-deliberative level, that they are accepted by the agent as a way of acting on her own reasons (the *auto-nudge* mechanism). These conditions rule out several forms of nudge paternalism that are accepted as permissible on the standard account offered by nudge advocates.

This account of autonomy-supporting nudge paternalism provides an answer to only one part of the thesis's overall question, however, because it concerns a specifically motivated subset of choice architecture, namely, that aimed (only) at the promotion of its target's well-being. There is another main motivation for PCA that requires our consideration if we are to provide a full answer to the thesis's question. This relates to the issue of modifying the behaviour of an individual *for the sake of others*. This second strand of analysis is the focus of the next chapter.

[98] I take it that these claims are supported by the fact that Moles's normative analysis of nudging includes autonomy-promotion as "justified within an anti-perfectionist framework" (2015: 645; see pp. 664-667 for the full relevant discussion).

5. Nudging as a Means of Moral Improvement

This chapter explores whether the political community can ever permissibly use choice architecture as a means of moral improvement and, if so, under what conditions. Its principal aim is to defend the view that there are a number of ways – including but by no means confined to those that aid the fulfilment of enforceable duties – in which choice architecture might permissibly be used to support the moral improvement of citizens. Systematic normative assessment of this moralising potential has been neglected in the literature. Thus, this chapter offers, to the best of my knowledge, the first full examination of nudging's moralising effects and the role that such effects might play within a theory of the political morality of public choice architecture (PCA).

It begins by showing how the normative analysis carried out in this chapter goes beyond the contours both of the standard accounts of nudging and of the existing debates over moral improvement (§5.1). After characterising moral improvement and considering how we might justify governmental efforts at bringing it about (§5.2), the chapter sets out an account of the conditions under which nudging for this purpose is morally permissible. This investigation is comprised of two parts. The first (and shorter) part examines the conditions under which it would be desirable to use moral nudges, of the kind deemed impermissible in the previous chapter, to secure conformity with enforceable duties (§5.3). The second (and main) part develops a detailed account of when moral nudges, principally of the kind deemed permissible in the previous chapter, might be used as a means of solving problems associated with nonenforceability. It sets out three kinds of interventions that are permissible, in part because they support agents in developing and exercising morally responsible agency (§5.4).

§5.1: Introducing Moral Nudges

Earlier in the thesis we drew an important distinction between two different kinds of motivation for using public nudges and their associated grounds of justification. In essence, the political community might seek to change an individual's behaviour either for the sake of improving her own well-being or for the sake of protecting the interests of others. The analysis carried out so far has focused on the former only. The motivation that interests us here concerns the use of nudge interventions to influence an individual's beliefs or behaviour in ways that aim to bring it into compliance or

conformity with moral reasons and norms – an aim I will label *moral improvement*.[99] As you would expect, this shift in the motivation for intervening activates a different set of normative concerns – or, more precisely, adds further considerations into the analysis that we have carried out already – which require careful, separate examination. In particular, efforts at moral improvement might be carried out for the sake of preventing harm to others (which raises questions about what is meant by harm-prevention), or they might be directed towards helping the target of these efforts to live well (which raises questions about the ways in which this might be so), or both (which raises questions about how to weigh the two considerations). This complex issue is the focus of the present chapter.

Before embarking on this inquiry, it is worth noting its novelty within the relevant philosophical literatures. Surprisingly, there has been very little discussion of how the permissibility of nudge interventions is affected by the different reasons that the state may have to intervene. That is, despite sustained *empirical* interest in whether and to what extent other-regarding nudges can bring about the desired behaviour change (e.g., increases in organ donation registration, charitable giving, pro-environmental behaviour, and so on), there has been very little discussion of whether and why we should employ these techniques, as distinct from the debate concerning their paternalistically-motivated counterpart.[100] This oversight is largely due, it would seem, to the fact that the most prominent nudge proponents, Sunstein and Thaler, have assumed a particular theory of political morality in their arguments which fails to discriminate satisfactorily between moral and nonmoral (i.e., prudential) reasons for intervention. According to their account nudges are justified when they increase social welfare, as determined by a more or less sophisticated cost-benefit analysis (Sunstein 2014a, 2016a). The first task of this section is to briefly outline this view and to explain why it is unable adequately to account for the normative concerns raised by the distinction between the self- and other-regarding considerations (§5.1.1). Seeing this opens up the possibility of a distinct category of PCA motivated not by improving

[99] There is a burgeoning debate in the literature about "moral enhancement". I elect not to use this term because it is now associated with a specific means, namely, moral bioenhancement (see §5.1.2 for more on this; Raus et al. 2014: 265). For this reason, the broader term, "moral improvement", is better able to capture a range of potential means with moralising effects without biasing the debate at all. (There remain, of course, important questions about what precisely constitutes moral improvement. These are addressed in §5.2.1.)

[100] Notable exceptions include Moles (2015) and Krishnamurthy (2015). The former offers an argument about when nudging is permissible within an anti-perfectionist framework, which includes using nudges for the purpose of increasing conformity with the principles of justice; the latter offers a more focused argument about whether we should employ nudges for the sake of global poverty alleviation.

social welfare as such, but by improving people's moral attitudes and conduct. Interestingly, this potential has been neglected in the debates over moral enhancement, as well as those about nudging. Hence, the section's second task is to briefly detail how the burgeoning literatures relating to moral improvement, most notably the debates over moral enhancement, have also overlooked the role that choice architecture might play (§5.1.2). The main aim of this section, then, is to highlight how this chapter both extends the nudge debate and brings it into innovative and productive contact with other burgeoning literatures.

§5.1.1: Moving Beyond Social Welfarism

As many have noted, including Sunstein and Thaler themselves (2003), several of the classic examples of nudges, to name a few – such as those related to tax fraud, organ donation, and pro-environmental behaviour – do not fit easily in the overarching normative policy programme of nudge paternalism into which they were originally embedded. This is because these interventions are not "motivated by the claim that *the person interfered with* will be better off or protected from harm", which is a central part of classic definitions of paternalistic intervention (Dworkin 2017; emphasis added).[101] Despite acknowledging this tension, nudge proponents have failed to offer any explicit account of how we should think about these kinds of cases. We can nonetheless reconstruct the type of account that underpins their views from the little they do say relating to their underlying theory of political morality.

On what grounds should we assess whether the political community is permitted to design PCA in ways that, say, discourage tax fraud or ensure that hiring decisions are not made on the basis of racial or sexist biases? Similarly, on what should we base our claims about whether states can permissibly nudge their citizens into registering as organ donors, giving to international aid charities to alleviate severe poverty, or reducing their carbon emissions and energy use to combat climate change? One view (and the view that we are reconstructing here) is that interventions of this kind are justified when – and because – they increase social welfare.

[101] As one commentator writes, "This ambiguity regarding just whose lives and interests are at stake in libertarian paternalism goes to the heart of Sunstein's project and raises deep questions about his proposal" (Kelly 2014: 180). Like Kelly, I think that this ambiguity represents a "fundamental lack of clarity" in the philosophical basis of the project. In their earliest work, Sunstein and Thaler very briefly distinguish "libertarian benevolence" from their main "libertarian paternalist" project, stating that it could be a complement to this project (2003: 1162). But this idea is not developed, or even mentioned in any detail, in their later co-authored work (Thaler and Sunstein 2009), or in Sunstein's more recent work (2014a, 2016a).

Clearly, people's actions sometimes have adverse third-party effects. On an economistic view, these negative "externalities" are considered to be a type of market failure (Bator 1958), namely, a situation in which an individual's pursuit of her self-interest leads to results that are inefficient from the societal point of view. Sunstein's more developed work relating to the permissibility of choice architecture makes explicit use of this market-failure framework (see, e.g., Sunstein 2014a: 16-17; Sunstein 2016a: 19-20, 31-32). It is as a result of this that he calls the category of nudges that are "designed to prevent people from harming others" *externality-reducing nudges* (2016a: 159). The central claim of this view is: If we take social welfare to be "the master concept" (Sunstein 2014a: 18) and accept that "it is perfectly legitimate for government to respond to market failures" (Sunstein 2016a: 32), then externality-reducing nudges should not be especially controversial in principle. The only issue to which we need to attend is whether or not such interventions would increase social welfare more effectively than alternative policy levers, according to a cost-benefit analysis.[102] We might call the justificatory framework for policies of this sort *nudge social welfarism* (see Korobkin 2009 for a similar view).

In the first instance, it is not clear why we should take an economistic approach to this political-philosophical issue regarding the permissibility of government intervention. But we might put this critique aside for now and try to engage with the underlying philosophical premises of the view. The thought seems to be a familiar one that it is permissible to interfere with a citizen's behaviour, and to do so against his will, in order to prevent harm to others (Mill 2005: 13) – where, in this case, the harmful effects on others are understood as those wrought by negative externalities. However, we need to recognise an important distinction relating to ways of understanding *harm*. Feinberg distinguishes between two notions of harm: (i) the broader, non-normative notion understands it in terms of setbacks to interests, while (ii) the narrower,

[102] It is for this reason that Sunstein focuses his attention on paternalistic cases, since these are the ones that he believes pose the hardest and most interesting normative questions (2016a: 54). The ultimate goal is the same in these cases: social welfare. In actions that affect only the person himself, social welfare is increased when his welfare is promoted. But, in these types of cases, considerations about welfare-promotion are more complicated than in cases of externality-reducing nudges, according to Sunstein, for two main reasons. First, an individual's well-being is promoted when it is steered in the direction that she would want to go, rather than in directions that reduce externalities. Second, whether or not an individual's well-being is promoted depends in part on the way in which the intervention affects her and whether or not it offends against her subjective desire for freedom of choice in that domain, since there will often be a welfare loss attached to manipulative means of welfare-promotion (though this is not necessarily decisive). In paternalistic cases, interventions are justified by the presence of a so-called "*behavioural* market failure" (Sunstein 2014a: 16 and 2016a: 19-20; emphasis added). Sunstein uses this analogy with market failure to identify the instances of cognitively-biased self-regarding behaviour that ought to concern us politically (Kelly 2014: 186).

normative notion views harm as a wrong (1987: 31-36). Nudge social welfarism, as we called it above, is based on the first notion; as with other consequentialist approaches, it thinks of harm as having to do with welfare (Woodward 1986: 818). The problem it faces, though, is that the harm principle is not directed towards prohibiting all kinds of harms to welfare interests – "only setbacks of interests that are wrongs" count as harms in the appropriate sense (Feinberg 1987: 36). Hence, justifying the permissibility of other-regarding nudges is not as simple as showing that the interventions would prevent people from harming others in the sense that they would prevent a setback to (individual third-party or social) welfare.[103] Social welfarism is not the right conception for dealing with other-regarding nudge proposals because: (a) it is sometimes not morally wrong to act in ways that diminish other people's welfare, and (b) there may be other kinds of wrong that do not refer to welfare losses.[104]

Rather, the fundamental category that we should be interested in is whether it is morally permissible to nudge people to do what they morally ought to do. If we return to the examples above with this in mind, we can see that what matters for the assessment of the nudge aimed at discouraging tax fraud, for example, is the fact that citizens are under a moral obligation to the political community to pay their fair share in taxes. These kinds of examples differ from those examined in the previous chapter in so far as they share a moral component. In other words, they all relate in some way to how we ought to respond to the claims that the moral status and rights of others make on us. The issue that grounds our evaluation of them is whether they can help us to better understand and fulfil our moral duties. It is moral goodness or rightness, not social welfare, which is the ultimate aim of these kinds of interventions. For rhetorical convenience, we can call the category of interventions that aim at moral improvement *moral nudges* or *moralising choice architecture*.[105]

---

[103] To do so, it would need to show that these setbacks to welfare are wrongs. A social welfarist may argue that we should specify the moral rightness and wrongness of acts in terms of their consequences and, specifically, consequences as they relate to the welfare of society. Yet no such arguments have been offered to this end (by, e.g., Sunstein or other nudge proponents). Nonetheless, even those who have explicitly defended social welfarist (or utilitarian) views – such as J.S. Mill – argue that harm to others' interests is a necessary, but not a sufficient, condition of justified interference.

[104] For an example of (b), see Slavny and Parr (2015) on cases of harmless, but wrongful, discrimination.

[105] The only explicit reference to moral nudges that I have found in the literature is in Van Ijzendoorn and Bakermans-Kranenburg (2011: 23), which, as we will see later (§5.2.3), offers an inadequate characterisation. Recently, Ishmaili M'hamdi et al. (2017) have introduced the concept of "other-regarding nudges" into the debate about healthcare interventions, specifically with regards to maternal choice during pregnancy, given the impact that this behaviour can have on perinatal morbidity and mortality and the potential development of chronic diseases later in life. In cases where this is done to prevent harm, this would count as what I am calling moral nudging.

The point here is only to show that, when we move beyond social welfarism as the grounds for assessing permissibility, we see that there is a specifically motivated subset of PCA that has not been adequately examined within the literature. The moral-improvement reasons for nudging have often been missed within the debate. There are a number of questions about how we might characterise and justify the aim of moral improvement that will have important implications for how we think about whether moral nudges are permitted for use by governments (these issues are addressed in §5.2). The reasons that support the goal of moral improvement do not necessarily say that we ought to pursue it by any means available. It may be the case that there are conclusive moral reasons to avoid certain means; or, more commonly, among means that are not conclusively ruled out, some means might be deemed to be better supported by moral reasons than others (Douglas 2014: 76). Our task here is to examine the permissibility of using PCA as a means to moral improvement – an assessment that will speak at times both to the evaluation of this type of intervention relative to other means and to the comparative assessment of different types of moralising choice architecture.

§5.1.2: Moving Beyond Current Debates on Moral Improvement

There are several ways in which we might try to improve an individual's moral conduct, from traditional legal and educative means to new, more controversial possibilities such as neuro-interventions and other forms of bioenhancement. As Douglas states, "There is […] much interesting work to be done in assessing the morality of different possible means to greater moral conformity" (2014: 76). Indeed in recent years the literatures relating to the permissibility conditions required for the use of moralising interventions, especially those that operate by directly manipulating our biology in various ways, have burgeoned. In light of this, it is surprising that so little work has been directed towards exploring the moralising potential of nudges, especially as they are grounded in similar neuroscientific discoveries and because it would appear that they could offer distinct advantages over other means. Before proceeding to that neglected task in the remainder of the chapter, it is worth briefly surveying the current debates on moral improvement in order to situate the discussion to come within its broader context and, in so doing, to highlight elements of its novelty within, and value for, the existing discussions about means of moral improvement.

(i) Biomedical Means

Much of the current debate is focused on the moralising potential of biotechnological means such as pharmaceuticals and electrostimulation of the brain. These are known collectively as *moral bioenhancement*. The central idea is that recent neuroscientific and technological advances may yield new means of "modifying morally significant aspects of motivation and behaviour" in ways that "make it more likely that [a person] will act morally, in some future period, than would have been the case if it were not used" (Savulescu, Douglas and Persson 2014: 91-92, 95).

Although the prospects of technically feasible bioenhancements lie in only the medium-term future, there are a number of studies that offer early signs of success with respect to the development of biomedical interventions that may be able to safely and reliably improve moral behaviour. This is due to the fact that scientific research is providing more robust evidence for biological correlates of morally-relevant traits. To take one example, neuroscientists have investigated the effects of serotonin, one of the main neurotransmitters in the brain, on moral behaviour and judgement. Specifically, they tested: (i) the difference in the willingness of serotonin-enhanced and non-enhanced participants to harm others via their responses to moral dilemmas similar in kind to the trolley problem; and (ii) the difference in these groups' evaluations of the fairness of monetary offers in the ultimatum game, and the associated rates of the rejection of unfair offers (Crockett et al. 2008, 2010a). The studies found that enhancing serotonin made participants both more likely to judge harmful actions as forbidden (though only in cases where harms were emotionally salient), and less likely to reject unfair offers.[106] These results have been taken to show that serotonin directly alters moral judgement and behaviour by increasing people's aversion to personally harming others (Crockett et al. 2010a).[107]

Empirical research of this sort, and its potential ethical and political implications, has recently generated significant philosophical interest. Different types of arguments have been advanced in defence of moral bioenhancement. A moderate view, offered by Douglas (2008) against what he calls the "bioconservative thesis", maintains that there are some reasons that support the permissibility of individuals morally

[106] Note that the second finding is not obviously preferable from a moral point of view.

[107] There are similar, but less robust types of findings relating to oxytocin and propranolol. With regards to the former, studies have shown that oxytocin (administered via a nasal spray) increases trust and cooperation, but this effect is limited to those people within one's social group (Kosfeld et al. 2005), while it reduces empathy for those outside of this group (De Dreu et al. 2010). With respect to the latter, propranolol has been found to reduce implicit racial biases (Terbeck et al. 2012). For a brief overview of the current state of the evidence, see Savulescu, Douglas, and Persson (2014: 93-94).

enhancing themselves via biomedical means.[108] A much more radical view, based on a thoroughly consequentialist account of morality, had argued that the development of biomedical moralising technologies should be prioritised and aggressively pursued by governments (Persson and Savulescu 2012).[109]

The main (interrelated) objections to moral bioenhancement are grounded in the importance of moral reasoning and the value of autonomy. Harris, for example, is in agreement about the need for moral improvement; but the problem with biomedical approaches, he argues, is that using neurally active drugs to directly manipulate people's emotions does not count as *moral* enhancement, properly understood: "tinkering with the emotions […] may make immoral behaviour less likely, but it does not enhance morality" (Harris 2013b: 171-172). The claim is that, although they may produce pro-social outcomes, some kinds of enhancements – specifically, those that employ *direct* means, in the way that biomedical drugs do – "fail to produce […] a deeper kind of moral improvement that is typically produced by traditional, deliberative enhancements" (Douglas 2014: 78, where he calls this 'the superficiality concern'; cf. Harris 2011). This first objection (from the importance of moral reasoning) draws our attention to the question of what should be classed as moral improvement, a foundational issue that I address in the following section (§5.2.1). It is strongly related to the objection from autonomy: in essence, the direct nature of biomedical means reduces the freedom to do wrong, and thereby undermines personal autonomy and the value that attaches to its exercise. These two issues – reasoning and autonomy – are clearly important features of the account of the political morality of nudging so far defended in this thesis. Precisely how the arguments made in the previous chapter translate into the moral domain is the issue that we will be engaging with fully over the course of this present chapter.

For now, it is worth noting three points about this debate that are relevant to the discussion about moral nudging. First, it opens up a sight of the potential contribution that moral nudges might make. The debate between advocates and critics

[108] Douglas' arguments have focused on interventions that aim to increase *one's own* moral attitudes and conduct, rather than on interventions that aim to increase the moral conduct of *others*, which is the focus here given our inquiry is centred on the governmental use of PCA to improve its citizens' moral attitudes and conduct (2008; 2014: 76 fn. 4).

[109] The view that moral enhancement is our most urgent task is grounded in a number of different claims: (i) the urgent global crises facing humanity (such as global warming and weapons of mass destruction), compounded with (ii) the biological limits to the human capacity for altruism, empathy, and a sense of justice, which are exacerbated further by (iii) huge changes to the conditions of living brought about by science and technology that have made us "less psychologically and morally fit for life" in this new environment (Persson and Savulescu 2011: 486). The adoption of moral bioenhancement in the near future is, they argue, necessary for the long-term survival of humankind.

of moral bioenhancement tends to set hypothetical future biomedical interventions that operate via direct emotion-modulating drugs against traditional, usually educational, interventions that aim at enhancing moral reasoning.[110] While the former are not currently available and are unlikely to be so in the near future, the oft-mentioned (non-normative) problem with the latter is that they have not proved to be adequately effective. Choice architecture, by contrast, is an institutional option that is available to us now and that may be more effective than, or at least an important complement to, traditional educational means (Zarpentine 2013). It therefore seems worth taking moral nudging seriously as a potential means of moral improvement.

Second, the debate over the ethics of moral bioenhancement can be read as taking on a somewhat similar structure to the discussion on the permissibility of nudging in the previous chapter: it seems that System 1 nudges share features with the direct modulation of bioenhancement, while System 2 nudges are intended as cognition-improving means and, hence, are closer to traditional moral enhancement. Third, in their interaction with this critique, defenders of biomedical means have suggested that their arguments on the permissibility of moral bioenhancement offer us "reasons to doubt that the [autonomy-based] objections will count decisively against nudge techniques" (Savulescu, Douglas, and Persson 2014: 109). This may be true; but, they provide no direct arguments to this end and this short shrift means that any morally relevant differences between the two means of influence are overlooked.[111] This gives us another reason, in addition to their current availability, to explore the moralising potential of PCA.

(ii) Traditional Means

There is also a wealth of ongoing and novel discussion in the literature relating to more traditional means. This comprises a more expansive set of debates than the discussion over bioenhancement, including issues such as: the cultivation of (liberal) virtues, the use of social norms in public policy, moral and civic education in schools, the role of the law in enforcing morality, and the idea that punishment can be used as a means of "moral fortification" (Howard 2017). The most relevant for our purposes are the

[110] Note that there are some examples that would seem to count as bioenhancements that we have known about for some time, e.g., that pupils who have sugary food and drinks are more likely to cause behavioural problems in the classroom, and schools often prohibit them (to some degree at least) for that reason and offer a different diet.

[111] This lack of attention from those within the moral enhancement debate rests, it appears, on the noncontroversial nature of moral nudges. But this cannot simply be assumed or intuited; and, to the best of my knowledge, no arguments have been given to this end.

arguments that relate to the use of social norms and the potential role of the legal system in supporting the development and exercise of citizens' moral reasoning and agency.

With respect to the former, there has been a debate over the implications of the psychological evidence for which (of the more traditional) political means of moral improvement should be favoured. McTernan has argued that, in light of the empirical thesis of *situationism* – which contends that human moral agency is "typically a function of the situation a person inhabits, or takes herself to inhabit, rather than any traits of character she putatively possesses" (Upton 2009: 104) – any kind of defence of the view that liberal virtues should be cultivated as "an effective means by which to make citizens behave" has been fatally undermined (McTernan 2013: 2).[112] She has argued that liberals ought instead to secure (or induce) stable patterns of behaviour from citizens via the use of social norms, i.e., "standard[s] of behaviour shared by a social group, commonly understood by its members as authoritative or obligatory for them" (Anderson 2000: 170). The central thought is that members of the social group will hold each other accountable to their norms, and that they will often "apply sanctions such as social exclusion or blame to those who fail to follow norms" (McTernan 2013: 11). This is of interest to the debate over nudging because some (moral) nudges work by making use of social norms. For example, choice architects sometimes design PCA that include "descriptive norms" relating to the levels of others' behaviour or "injunctive norms" relating to the level of others' disapproval (Cialdini et al. 2006: 3), for the sake of influencing people's other-regarding behaviour.

McTernan's argument, however, seems problematically to assert that social norms – and, by extension, the nudges that utilise social norms – are mutually exclusive from the deeper kinds of moral learning and the resultant behaviour change that accompanies, and is the primary aim of, civic education (which, for Callan, is the same as "the cultivation of civic virtue") (Callan 2015: 491-492). On her view, this is due to the psychological evidence about how our choice environments affect our moral agency. But it is not clear why this would necessarily be the case. In Callan's response to this argument, he states that:

> "the inference to draw from the experimental evidence is not that virtue does not exist but that its rationally expected exercise is typically indexed to a range of circumstances within which the agent has been habituated to act and feel as she should, and where

[112] For more on the (philosophical) situationist challenge, see Doris (2002) and Harman (1999).

temptation comes from predictable sources she has been taught to manage well." (Callan 2015: 495)

This seems right in so far as it goes (for more on this and related issues, see Kristjánsson 2000 and Adams 2006: 114-158); but it offers only a negative defence of traditional civic education, rather than any active engagement with the psychological evidence, or a positive account of how traditional educational means might need to be updated or supplemented in light of this evidence. This reveals the opportunity for exploring the potential of moral choice architecture – which by its very nature takes situationism seriously – as a way of achieving the goals of moral and civic education, and potentially doing so more effectively than some current policies that are insensitive (or less sensitive) to moral psychological research.

There has been some related discussion about the role of the legal system in recent years. I will briefly make reference to three novel arguments. First, Shiffrin has argued that "legal standards", i.e., directives that incorporate thick, substantive terms that require the direct application of the background principle to a fact situation, may have the virtue of inducing greater levels of moral deliberation, relative to "legal rules", i.e., directives that instruct people to respond in a determinate way to the presence of delimited triggering facts (Shiffrin 2010: 1214-1215). Second, Brownlee and Child (2012) have argued that there is a case for the law being able, in principle, to guide us morally through (i) its instructions, (ii) its examples, and (iii) its motivational prompts. Third, relatedly, Brownlee has presented a defence of the view that the law might offer a means of cultivating virtue through its "setting a moral example that we have good reason to emulate" (Brownlee 2015: 2).[113] Moral choice architecture may also be able to play similar roles. If so, there may be interesting comparative assessments about the conditions under which nudging may offer advantages over the legal measures. This provide further reason to take it seriously as a potential means of improving people's moral capacities and conduct.

§5.2: Moral Improvement

So far we have established that there is interesting work to be done relating to moral nudges and the permissibility conditions for their use by governments. There are several complex questions and issues that this inquiry will need to address. The two central

[113] For a discussion of whether and how nudges can be used to in the policy-led process of cultivating virtues, see Niker (forthcoming).

issues that we need to get clearer on first are: (i) what we might mean by "moral improvement" (§5.2.1), and (ii) the reasons that the political community may have to try to promote moral improvement (§5.2.2).

§5.2.1: Characterising Moral Improvement

What constitutes moral improvement? We can begin by distinguishing between an individual's *behaviour*, on the one hand, and her *beliefs* or *motivations*, on the other. Accordingly, we can separate analytically three distinct elements of the "anatomy" of moral improvement: (i) behavioural improvement, (ii) improved insight, and (iii) motivational improvement (DeGrazia 2013: 362-363). As we will see below, the second and third elements may overlap in some instances, but they can come apart too.

The first element, improvement of moral behaviour, is outcome-oriented; that is, it is concerned with increasing the occurrence of right action by bringing about greater conformity with what a person has moral reason to do. In these cases, all that we would mean by nudging a person for moral improvement purposes is that the intervention improves her morally relevant behaviour in some way. Of course, behavioural improvement may occur as a result of improved insight or motivational improvement – indeed this may be considered a part of the (instrumental) value of these two elements. But it can occur independently, too, such as in cases where someone is forced to perform the right action. It is these cases that this category is principally intended to capture (as distinct from the other elements).

The second and third elements concern the beliefs and motivations that explain our behaviour. Let's begin with the second: improved insight. Unsurprisingly, our beliefs about what we have reason to do direct much of our behaviour. We often behave in a particular way – say, one that prevents harm to the environment – because we hold a particular belief or set of beliefs, in this case, the belief that we ought not to contribute to climate change (which could be grounded in the impersonal value of the natural environment or in intergenerational duties, or both). Independent of the beliefs that we actually hold, morality gives us reasons to act, and not to act, in certain ways; and our beliefs can correspond to a lesser or greater extent to these moral reasons. Morality sometimes also gives us reasons (not) to act from certain beliefs. Hence, a person can fail to act in accordance with morality either in virtue of being unaware or ignorant of relevant moral reasons or because she holds irrelevant or mistaken moral beliefs. The second kind of moral improvement, then, refers to the improvement of an individual's moral insight, such that her beliefs and judgements about what she ought to

do are more in line with the moral reasons that there are. This epistemic dimension is what moral education, for instance, takes as its primary goal. As already mentioned, improved insight will lead to behavioural improvement in many cases. This is because people can be motivated directly by their beliefs about reasons to act in accordance with them. This highlights that moral insight and moral motivation are not as separable as the taxonomy above may have initially suggested.

There is an important and highly contested debate in philosophy, and a related empirical investigation in psychology, relating to *moral motivation*.[114] It is beyond the scope of the present inquiry to set out an account of the nature of the connection between moral judgement and motivation; however, it will be helpful to offer a few indicative comments about some issues relating to moral psychology. The first relates to why I take it to be necessary to include motivational improvement as a separate category in our analysis. The second, by contrast, returns to the idea that moral insight and motivation are not completely separable, by taking a view on the role that emotions can play in informing us about what is right.

One of the main strands of the debate over moral motivation revolves around the issue of whether moral beliefs and judgements motivate necessarily, as the *internalism thesis* contends, or whether they do so only contingently, as the *externalism thesis* holds (Rosati 2016).[115] Korsgaard has distinguished these two theories thus:

> "An internalist theory is a theory according to which the knowledge (or the truth or acceptance) of a moral judgement implies the existence of a motive (not necessarily overriding) for acting on that judgement. If I judge that some action is right, it is implied that I have, and acknowledge, some motive or reason for performing that action. It is part of the sense of the judgement that a motive is present: if someone agrees that an action is right, but cannot see any motive or reason for doing it, we must suppose, according to these views, that she does not quite know what she means when she agrees that the action is right. On an externalist theory, by contrast, such a conjunction of moral comprehension and total unmotivatedness is perfectly possible: knowledge is one thing and motivation another." (Korsgaard 1986: 8-9)

[114] There is an extensive philosophical literature on moral motivation. For a helpful overview of the debate, see Rosati (2016).

[115] A second strand in this debate is whether moral beliefs require the presence of a desire or another conative state in order to motivate us to action, as Humean views argue (Smith 1987; Railton 1986b), or whether beliefs about moral reasons motivate on their own, as anti-Humean views maintain (Nagel 1970; Scanlon 1998). I do not take sides on this issue, especially since I believe that the arguments I defend are compatible with both accounts of human action.

.

A historical survey of the internalism/externalism debate shows, nonetheless, that the original distinctiveness of these two positions has been eroded over the years, as versions of internalism have become "increasingly qualified" in order to retain their plausibility against externalist critiques (Rosati 2016).[116] We have already recognised that people can be directly motivated by their beliefs, but this leaves it open whether this is a necessary or contingent connection. The empirical evidence at the heart of this thesis seems to support the latter view. We are interested in certain problematic cases of automaticity, in which an individual acts in a way that is not (necessarily) motivated by her beliefs – either when she acts in a certain way without believing that there is a reason to do so, or when she fails to act in a certain way despite believing that there is a decisive reason to do so. One prominent example is that of implicit biases in which an individual who holds the moral belief that racist or sexist judgements are wrong, and therefore does not believe that there is a reason to discriminate along racial or gendered grounds, nonetheless acts in a prejudiced manner on account of (relatively) unconscious and automatic cognitive processes (Brownstein 2015). Thus, the cognitive and behavioural science evidence highlights that, even though the connection between moral beliefs and motivation might be strong and predictable, it is ultimately contingent. This suggests that there may be motivational improvement strategies aside from that which follows as a result of improved moral insight.

Moral beliefs, then, can sometimes issue in moral motivation, but this is not to say that the relationship between motivational states and moral insight is unidirectional. The second comment we need to make about moral psychology relates to the idea that certain interventions that might be described as seeking to bring about motivational improvements in so far as they influence affective (or, more precisely, affective-cum-motivational) states, such as emotions, may have an important role to play in moral insight. There are several philosophical theories of the emotions that offer competing accounts of the relationship between emotion, value, and reason (see Deonna and Teroni 2012 for an overview). My own view aligns with what I take to be the most common in the current literature, which is generally referred to as the *perceptual theory* (see De Sousa 1997: 149-158 for a discussion of the analogy with perception).[117] This

[116] For instance, Shafer-Landau treats the defeasibility of moral motivation under certain conditions – such as severe exhaustion, serious depression, or overwhelming contrary impulses – as supporting externalism (2003: 147-148), while Smith (1994) treats similar defeasibility conditions as being compatible with a form of internalism.

[117] The attractiveness of this theory rests in its steering a middle course between two opposed accounts of emotions – the feeling theory and the judgemental theory – which each have some plausibility, but which

name is somewhat misleading as there are many different perceptual theories of the emotions (e.g., Tappolet 2016; Roberts 2013; McDowell 1998; Deonna 2006; Zagzebski 2003; Nussbaum 2001; Brady 2013); but the central thought is that, in their capacity as analogous to perceptual states, emotions have an epistemic function in moral cognition. That is, emotions can play a role in informing us about our moral reasons.[118] The least controversial claim to this end, and one that is heavily supported by the empirical psychological evidence (Damasio 1994), is that emotions are detectors of salience: they draw our attention to morally relevant features of the situation and, in so doing, can help us better to recognise our (moral) reasons for action (De Sousa 1997). Classic examples include anger at the sight of an injustice and guilt or shame when one realises that they have violated a moral standard.

§5.2.2: Justifying Moral Improvement

With respect to the second issue, that of justifying moral improvement as a legitimate aim, the key question is: Why might the political community wish to improve an individual's moral conduct (in one or more of these three ways)? This is a derivative of the moral general question "Why be moral?" and, as such, requires that we give an account of why we and others have compelling reason to be moral (Scanlon 1998: 148; see also Dworkin 2011: 191-199).

There are two possible reasons for designing PCA for the purpose of moral improvement, both of which are grounded in its instrumental value. The first is to protect the interests of others, which would be (wrongfully) harmed if the moral improvement is not brought about. The second is to support morally responsible agency, which is a source of ethical value for the target of the intervention. The second type of reason is more controversial than the first. In my view, the question of when it is permissible to use moral nudges needs to assess both kinds of reasons (these normative analyses are carried out in §5.3 and §5.4, respectively), as well as to examine the issue of how to weigh them against each other in cases of potential conflict (§5.4.4).

Governments have always engaged in efforts to improve citizens' moral conduct. The most common reason for doing so is in order to prevent an individual from acting in ways that harm others. Although the ideal of autonomy and its value to a human life morally constrains the use of public nudges and other interventions in

---

are both "ultimately unsatisfactory" (Tappolet 2014: 168). For more on this, see Deonna and Teroni (2012) and Tappolet (2014, 2016).

[118] For what it's worth, I favour an "evaluative perception" account of this, such as the neo-Aristotelian theory offered by McDowell (1998) and Wiggins (1987).

various ways, as we saw in the previous chapter, liberals are in agreement that some violations of an individual's autonomy are permissible. The interests of others can be sufficiently weighty so as to justify the claim that we have *moral duties* to attend to them in an appropriate way, either by doing something (in the case of positive duties) or by refraining from doing something (in the cases of negative duties). Many of these duties are, in principle at least, enforceable by the political community (more on this in §5.3). In these cases, governments themselves may be under a moral duty to protect their citizens from certain kinds of harm. When considering this reason *in isolation*, it does not seem particularly relevant what form the moral improvement takes so long as the outcome is secured (unless, as is sometimes the case, the duty in question requires that one fulfil it from a certain motive).

The second justification of moral improvement relates to the promotion of the well-being of the target of the intervention. In my view, a person's life goes better when she recognises for herself and appropriately responds to the moral norms and reasons that apply to her in a situation, than when she is forced to perform the actions that are required by these norms and reasons. Consider two citizens, Amy and Amin, both of whom act in the ways that morality requires over the course of their adult lives. Amy behaves in these ways because she is the target of some sort of mind control, perhaps psychologically manipulative nudges, that get her to perform the actions automatically. By contrast, Amin acts in these same ways, but does so because he is receptive and reactive to the moral reasons that exist. A simple way of capturing the difference between Amy and Amin is to adopt the distinction between conformity and compliance with moral reasons.

According to Raz, we *conform* with a reason for action when we perform the action that the reason favours, while we *comply* with it when we perform the action in direct response to the reason for its favouring of that action (1999: 178-179). To illustrate this contrast, imagine a case in which Adam has a decisive reason to stay at home with his partner Eve because Eve has just received news of a friend's death and Adam's presence around the house would provide her with comfort and support.[119] If Adam stays at home, then he conforms with this reason. But, in cases of conformity, he may have stayed at home not out of an appreciation of the reason provided by Eve's need for moral support; rather, what motivated this action is that he wants to be close to his computer in case he gets some inspiration about how to finish the current chapter

[119] This case is a lightly revised version of the case Raz uses to illustrate this distinction (1999: 178-179).

of his novel about which he is currently suffering from writer's block. By contrast, if Adam not only stays at home, but does so (a) because he recognises that this will bring Eve some comfort at this difficult time, and (b) is motivated to do so for this reason, then he complies with the reason. We can assume that Adam is acting autonomously in both cases, since he might be equally responsive to his own reasoning. There is, nonetheless, an important difference between conformity and compliance, in my view, at least from the point of view of their *ethical value*. What explains this?

Let's return to the example of Amy and Amin. Amy's life is one of moral conformity, while Amin's is a life of moral compliance. On my view, Amin's life includes an important aspect of what it means to live well, which Amy's life lacks, namely, *moral responsibility*. Part of this value comes from the fact that morally responsible agency often tends to be autonomous agency. In Amin's life, for instance, his actions result from his being responsive to his own reasoning, such that he can be described as the author of these actions in the relevant sense. The "Autonomy View" of moral responsibility, which claims that "beings are responsible just in so far as they are autonomous" (Wolf 2008: 261), does capture the (true) idea that the reason why Amy is not morally responsible is because her actions are the result of someone manipulating her to act as she does: if Amy has no control over these actions, it is easy to see why she cannot be said to be responsible for them. But this view fails to capture the whole picture. Even if one holds that morally responsible agency needs to meet this "control condition" to the same degree as autonomous agency (a view which is often contested, see, e.g., McKenna 2008 and Wolf 2008), there remains a difference between the kinds of "epistemic condition" carried by two types of agency.

As we saw in the previous chapter, what matters for autonomy is that an individual is able to be responsive to her own reasoning; hence, any epistemic condition that there might be for autonomous agency would be restricted to matters relevant to the goals, attitudes, beliefs, values, and so on that serve as the basis for her self-authorship. A stronger epistemic condition is required for moral responsibility, however. On the view that I favour, Amy is not morally responsible *not only* because her actions are controlled by someone else (i.e., because she is not autonomous with respect to her actions), but because this manipulation denies her "the freedom to see things aright" (Wolf 2008: 273). Amin, by contrast, is morally responsible not only because he acts autonomously, but because his embracing and acting on relevant moral values is an expression of his understanding and appreciation of the reasons that exist in favour of

acting in accordance with these values. This process of appreciating moral reasons, like the process of acting autonomously, cannot and need not be done independently of external influences. The public environment (or what Hurley calls the "public ecology") can, and should, play a role in helping us with *both* the control (autonomy-supporting) condition, as well as the content-sensitive epistemic condition, required in the development and exercise of morally responsible agency. Moral choice architecture is one way in which this can occur. As Susan Wolf explains:

> "It is by being rationally persuaded that these values are good ones that the agent makes them her own in a way for which she is responsible. But there is no analogous story to be told of the agent who acquires bad values from his culture. We cannot say that the racist is responsible for his racism if it results from his understanding about what is good about racism – for there *is* nothing good about racism for him to understand. Nor can we say that the racist is responsible for his racism if it results from his understanding about what is bad about racism – for no sane person chooses values because he understands them to be bad." (Wolf 2008: 269-270)

On this so-called "Reason View", therefore, a person is morally responsible for her actions "if she was not only free to govern her actions in accord with her values, but that she was able to revise her values in accord with reason and truth" (Taylor 2008: 269).

Consequently, moral compliance, and the moral responsibility that attaches to it, can be understood as *responsiveness-to-reasoning plus*, if you will, where the "plus" refers to the addition of the more stringent epistemic condition required when we bring in moral values. Acting autonomously is valuable, but this value has an epistemic conditionality in place which requires that the agency is responsive to the reasons that there are. In the case of Adam, moral compliance requires that he recognises the reason there is to support Eve (i.e., (a) above, which can be described as "reasons-recognition"), and that an appreciation of this reason is what motivates his conduct (i.e., (b) above: "reasons-reactivity") (Fischer and Ravizza 1998). Similarly, Amin exercises morally responsible agency because, by stipulation, he recognises and reacts appropriately to moral reasons. This second justification for moral improvement maintains that, on account of the ethical value to the agent herself of morally responsible agency, there is a *pro tanto* reason in favour of adopting means of moral improvement that support compliance over those that merely secure conformity.

§5.2.3: A Taxonomy of Potential Moral Nudges

We have distinguished between three types of moral improvement – improvements to behaviour, motivation, and insight – and two kinds of reason that may be used to justify these moralising interventions – to prevent harm to others and for the ethical benefit of the agent. Our task now is to examine whether moral nudges can permissibly be used as a means of moral improvement and, if so, under what conditions. To give some structure to this analysis, it is instructive to clarify the different ways in which PCA might interact with the characterisations and justifications of moral improvement.

There are three kinds of relevant considerations. The first is that morality gives us reasons to act in certain ways, as well as reasons to refrain from acting in others. In many of these cases, these *duty*-based reasons in themselves require only that we act in conformity with them. Hence, behavioural improvement would be sufficient in such cases. Nonetheless, and this is the second point, sometimes morality gives us reasons to do (or to refrain from doing) things from certain *motives*; that is, we are sometimes under moral duties that require compliance. In these cases, behavioural improvement would not be enough; it must be accompanied by, or more precisely follow from, relevant moral beliefs or motives. The third consideration is that, in addition to the duty-based reasons outlined in the first and second points, which are justified via the prevention of harm to others, there is also an *ethical* consideration that favours the use of interventions that aim at changing a person's behaviour by improving her moral insight and/or motivation, wherever feasible.

As a result, it seems that we can distinguish between four different potential aims of moral nudging:

(1) to change an individual's other-regarding behaviour, regardless of her beliefs;
(2) to help bring an individual's behaviour in line with her existing moral beliefs, though passively on the part of the agent;
(3) to induce deliberation in order to support an individual's moral reasoning, which may:
    (a) support her in bringing her own behaviour in line with her existing moral beliefs; or.
    (b) bring about changes in her behaviour and sometimes her beliefs through a process of better recognising moral reasons; or,
(4) to change an individual's beliefs to be more in line with the what morality requires.

The distinction between "System 1" nudges and "System 2" nudges that played a role in our prudential analysis is relevant here, too. Recall that the former bring about behavioural modifications via (relatively) automatic processes, while the latter act on us in ways that encourage some conscious engagement. Both (1) and (2) will be delivered via System 1 nudges; while (3) and (4) require System 2 nudges. With respect to (1), the moral nudge would aim to induce *behavioural improvement* for the sake of preventing harm to others. This type of intervention mirrors the psychologically manipulative System 1 nudges delineated in the previous chapter. Indeed, some describe the category of moral nudging in general in these terms: e.g., Van Ijzendoorn and Bakermans-Kranenburg define moral choice architecture as: "a myriad of situational manipulations that effectively change human behaviour in [morally] desirable directions *without* changing their moral reasoning, dispositions or motivations" (2009: 23; emphasis added). This view seems importantly incomplete, given the other three categories of potential moral nudges. The purpose of (2) would be to use moral choice architecture as a means of securing *motivational improvement*. This category of System 1 nudge parallels the auto-nudge mechanism set out in Chapter 4. Like (2), (3a) is a form of motivational improvement, but it differs in so far as it seeks to bring this end about via deliberative processes, much like those involved in the nudge-autonomy mechanism. Both (3b) and (4) aim at encouraging *improved insight*. Relating to the discussion above about morally responsible agency, these two possible types of moral nudging seek to persuade citizens about the reasons that there are to act in particular other-regarding ways.

On my view, the project of moral improvement via PCA should be primarily directed towards supporting and stimulating citizens' moral agency so that they are better able correctly to identify and respond appropriately to moral reasons, and thereby bring about the desired or required behavioural improvements in an autonomous and morally responsible way (i.e., (2)-(4), in various ways). This is because PCA is *doubly justified* in these kinds of cases: it protects others within the political community from harm, and does this in a way that supports the agent in living well. Nevertheless, the well-being-based justification is, as we mentioned above, a more controversial type of justification than the harm-prevention-based one. We might accept, for example, that it would be *better* if people gained moral insight and motivation with respect to fulfilling their duties – that this would make their lives go better – but nonetheless worry about how much conformity, if any at all, we should sacrifice in our attempts to secure moral compliance. I address this complex issue later in the chapter (see §5.4.4).

The remainder of the chapter explores the complexities involved in assessing the ethics of moral nudging, which it does by carrying out two strands of normative analysis. The first, in the next section, leaves aside considerations relating to the ethical value of moral responsibility to explore the permissibility conditions for the category of nudges that aims to secure behavioural improvement, i.e., (1). It asks when, if ever, we have good reasons to use (manipulative forms of) choice architecture, as compared with more traditional coercive means, for the purpose of securing conformity with our moral duties. The second part, carried out in §5.4, examines the conditions under which it is permissible to use PCA as a means of supporting morally responsible agency.

§5.3: Moral Duties, Enforceability, and Choice Architecture

It follows intuitively from the idea that the political community is entitled to force its citizens to perform a given action that it would also be entitled to act on its citizens in a noncoercive way in an attempt to get them to perform this same action (assuming the same effectiveness). One claim that we might make about moral nudging therefore is that, when an individual has an *enforceable duty* (not) to act in a particular way, we are permitted to use PCA in any way that may bring about the morally required behavioural improvement. The reference to 'in any way' in this statement refers, most particularly, to the distinction that we drew in the previous chapter between manipulative and non-manipulative types of nudges. In the prudential realm, we have seen that only the latter are morally permitted. This analysis does not neatly extend into the moral realm, as we shall see.

The moral constraint in the earlier cases was provided by the ideal of autonomy. Yet, despite its weighty value, autonomy is but one value among many, and there will be occasions when state support of these other values will override its commitment to personal autonomy. When it comes to citizens' other-regarding behaviour, the idea that sometimes violations of autonomy are morally permissible is generally considered to be uncontroversial. Liberals typically hold that the central justification for infringing on a person's autonomy is the prevention of harm to others.[120] We are under several duties to others that are, in principle, enforceable. Thus, the political community has a moral permission – and, in some cases, an obligation – to ensure that people perform these duties. Following Tadros (2011), we can call this the *duty view* and, for our purposes, it

[120] One way of cashing out this thought is that a person's autonomy is not sacrificed in these cases, because he or she has no claim, grounded in autonomy, to have access to options to wrongfully harm others (Wall 2013: 105).

can be summarized thus: in cases in which an individual has an in-principle enforceable duty to serve a particular end, intervening in that individual's behaviour to get her to serve that end does not wrong her.

We now need to say a bit more about the issue of enforceability. Enforceable duties are duties that it is permissible to coerce – i.e., threaten or force – a person into fulfilling. These are distinguished from nonenforceable duties, which can take one of two forms. The first concerns duties that, despite being enforceable in principle, it is not morally desirable for the state to coerce people into performing, such as duties to aid others. The state may often permissibly use other policy levers, such as persuasive techniques, as a means of getting citizens to perform these duties, but it cannot force their fulfilment. The second concerns duties that are nonenforceable in principle. Some duties, by their very nature, require that they are performed freely, such as the duties of benevolence or friendship. There is a range of factors that bear on whether a duty is enforceable or not, which include (but are not exhausted by) the following set of considerations: "the moral significance of the duty, the extent to which it is likely that the ambitions that ground the duty are likely to be advanced through enforcement, the harms that will typically be imposed on the duty bearer through enforcement, [and] the extent to which it is important that the person acts on the duty for good reason rather than because she is forced to do so" (Tadros 2011: 132).

The most innovative uses for moral choice architecture, in my view, involve the ways in which it can help the political community in overcoming some of the problems associated with nonenforceability, which we will explore in the next section. The focus of this section is on the use of moral nudges in cases where it is permissible to coerce people into moral conformity. In such cases, is it also permissible to use nudges – in particular, the psychologically manipulative System 1 nudges that we earlier ruled out as impermissible (when used for prudential ends) – as a means of eliciting behavioural improvement? And, if so, what are the reasons either in favour of or against using nudges rather than (additional) coercive means for this purpose?

Recall that psychological manipulation occurs when an individual is influenced by causing changes in mental processes other than those involved in understanding (Faden and Beauchamp 1986: 366). As we have defined them, System 1 nudges are manipulative, when they are not consented to, because they act on their target by directly influencing their automatic cognitive processes, thereby bringing about behavioural modifications passively from the point of view of the agent being acted on.

It seems possible that, in line with what the duty view claims, we could use these nudges in cases where citizens have an enforceable duty to act in a particular way. But when might it be desirable to do so? In one type of case, this question is the same as: when might nudging be preferable to (further) coercion? This is the type of case we will look at first.

The most obvious answer to this question would appear to be grounded in the non-moral value of *efficacy*. Nudges, even psychologically manipulative ones, are permitted in cases where they are more effective than coercive measures at increasing moral conformity (Moles 2015: 660). There are two ways in which this might be the case: either they are just as effective as coercive means, but are cheaper and easier to implement (and therefore more efficient) or, more simply, they are better at achieving the intended behavioural results (perhaps because they operate covertly). For example, research has found that behaviourally informed design interventions may have a significant effect on reducing tax evasion. Specifically, changing the design of self-report forms by moving the declaration-of-honesty signature box from the end to the beginning of the form has been shown to reduce instances of fraudulent reporting.[121] The important subtle change relates to temporality. If the signature is given at the end, a person may have already provided a dishonest self-report, and we know that "immediately after lying, individuals quickly engage in various mental justifications, reinterpretations, and other 'tricks' such as suppressing thoughts about their moral standards that allow them to maintain a positive self-image despite having lied" (Shu et al. 2012: 15197; see also Bandura et al. 1996). There may be different ways of interpreting how this particular nudge acts on its target; but, for our purposes, let us assume that it works as a form of unconscious moral priming.[122] Let us also assume that

[121] Instances of cheating (by overclaiming income) in this laboratory experiment was 79% in the condition where the signature box was at the bottom and only 37% when it was placed at the top of the page, relative to 64% where no signature was asked for (Shu et al. 2012).

[122] It may be the case that this effect on reporting behaviour is explained by reference to the nudge acting to "make morality salient" to the person right before they act (since, e.g., signatures are one possible way of activating attention to the self) (Shu et al. 2012: 15198). And this could, in turn, be thought to work as a form of unconscious moral priming or rather by supporting our moral reasoning when we are deciding what figures to report, i.e., by making moral reasons more salient so that they are less likely to be overpowered by our self-interested ones (see, e.g., Batson and Thompson 2001 for more on this phenomenon). Here, we assume the former, which is also in line with the view taken by the authors of this study. There is an increasing amount of empirical evidence relating to the kinds of potential means that could be used for unconscious moral nudging, so we could reasonably expect that there would be other possible moral System 1 nudges that could have a similar effect with respect to discouraging fraud. For instance, researchers at Harvard have found that displaying cues such as moral quotations at the bottom of emails and pictures of moral leaders such as Mahatma Gandhi can trigger implicit (or unconscious) psychological processes such that people feel discouraged from behaving immorally (Desai and Gino, unpublished). Similarly, it has been shown that cues relating to childhood, such as recollecting

it is more effective, in one of the two ways specified above, than increased penalties – an assumption that is supported, in part at least, by the UK Behavioural Insights Team's promotion of this policy (Cabinet Office 2012: 14-16). Despite there being different views about the scope of justified coercive taxation, there is a general consensus in both political philosophy and practice that each citizen has an enforceable duty to pay whatever taxes they are deemed to owe to the political community. Hence, on this duty view, it would be permissible for the state to use these System 1 nudge policies to secure moral conformity.

But this view, as it stands, does not represent the whole story, because there may be morally relevant differences between these nudges and standard coercive measures, such as (threats of) fines and imprisonment. One way of addressing this is to explore the question about whether manipulative nudges might ever be described as a means of *enforcement*. Given that choice architecture is characterised as being the noncoercive alternative to these traditional measures, this question might seem misguided: there is no threat of force and no legal consequences with respect to the criminal law if someone opts out; hence, nudges cannot be thought of as an enforcement mechanism. This is true; but such a view seems too simplistic in the case of psychologically manipulative choice architecture. We can see this by referring to a more detailed definition of coercion, which recognises a distinction between two kinds:

> "We typically say that *A* coerces *B* to do *X* when *A* gets *B* to do something by threatening to harm *B* or by making *B* worse off in case *B* should not do *X*. We also often say that coercion interferes with one's freedom or autonomy, that if *B* is coerced into doing *X* (or does *X* under duress), then *B*'s action is involuntary." (Wertheimer 2015: 2)

This type of distinction has been drawn as far back as Aristotle's *Nicomachean Ethics* (2009: 38; at 1110b 1-7). There, Aristotle more or less explicitly distinguishes between two types of actions that are enforced (*biaia*). One kind is enforced because the agent contributes nothing to the act, thereby making them unintentional (*akousia*) or "involuntary", in the terms used above. A second kind are those acts where the agent does contribute something, but does so under threat; these are intentional (*hekousia*) or "chosen, viz. as preferable to not performing the act with the consequences that this would have" (Engberg-Pedersen 1983: 182). Importantly for our purposes, manipulative

---

memories of your first experience of riding a bike, can result in less cheating behaviour and increased generosity toward others (Gino and Desai 2011).

nudges can be thought of as being a form of *coercion-as-involuntariness*, as we might call it, while traditional coercive means operate as *coercion-as-threat*.

It is this, I think, that explains the charge that, even when violations of autonomy are permissible, manipulative System 1 nudges can nonetheless be morally problematic on account of their insidiousness. Hausman and Welch suggest this type of view when they highlight that "there may be something more insidious about shaping choices than about open constraint" (2010: 130). Essentially, the thought is that the insidiousness associated with the mechanism by which manipulative nudges operate could be a wrong-making feature that renders the practice an impermissible means of serving the permissible end of enforcing morality.

There are two comparative elements to this claim. The first is that nudges of this sort bypass our reasoning process in a way that we might think of as being akin to mind control, and this is more insidious than using explicit coercion to exercise control over the actions people choose. The thought is that psychologically manipulative nudges may disable citizens from being able to recognise and endorse the reasons to perform their duty, which is something that remains open to those who are subject to coercion-as-threat. The latter policies are an example of what Clayton and Moles (forthcoming) call "Frankfurt interventions", namely, interventions that protect the interests of others while ensuring an opportunity for the agent to perform the required action freely.[123] Standard legal prohibitions allow an individual to act on her own moral reasoning free from interference so long as her actions do not violate any of her enforceable moral duties.[124] By contrast, psychologically manipulative nudges act on everyone and, as a result, they manifest disrespect in citizens' ability to form and act on moral beliefs that adequately track the moral reasons there are.

The second comparative element relates to the fact that, in some cases, part of the justification of the coercive policies comes from their communicative effects. Coercion plays the important role of communicating the political community's judgements about a given kind of conduct, and it is this expressive value that sometimes grounds part of its permissibility (Duff 2001). This communicative potential, and its

[123] These are named in reference to part of Frankfurt's argument in 'Alternative Possibilities and Moral Responsibility' (1969).

[124] The problem in the case of real-world examples like legal coercive measures is that, unlike the fictitious Frankfurt interventions, they are not able to kick in to prevent a person acting against her enforceable duties. Consequently, they cannot always prevent harm to others (but only to punish the agent afterwards). Hence, the preventative nature of mind control measures does provide a *pro tanto* reason in their favour (for more on this, and the issue of the right to do wrong more generally, see Savulescu, Douglas, and Persson 2014: 100-108).

associated potential as a means of supporting improved moral insight and motivation, is not open to the kinds of nudges that we are discussing in this section. These two points highlight that manipulative nudges block the possibility of developing and exercising morally responsible agency in a way that coercive measures do not and, as a result, deny citizens the opportunity to access the ethical value that attaches to the exercise of such agency.

Hence, it might sometimes be morally impermissible (or at least, undesirable) to use moral nudges as a means of securing moral conformity, even though it would be permissible to force people to conform. But this concern requires only that we qualify, not abandon, the duty view. The two considerations outlined above offer *pro tanto* (or defeasible) reasons; and so it is possible that there will be some cases in which the political community will have decisive reasons to use these kinds of moral nudges rather than (further) coercion. This may be the case if: (a) nudges were (much) more likely to be successful, (b) nudges were (much) less costly to implement (on the assumption of similar efficacy), or (c) nudges would not (but coercion would) impose on the target burdens greater than that for which he or she is liable.[125] When one or more of these defeasibility conditions is met, psychologically manipulative nudges are all things considered morally permissible, despite the fact that their insidiousness does make them worse than coercion in one way.

§5.4: Moral Responsibility, Living Well, and Choice Architecture

This is not the only way in which PCA might be used to bring about moral outcomes. We distinguished four distinct potential aims of moral nudging in §5.2.3, and have so far focused only on the first. In this section, our focus shifts to the question of whether choice architecture can play a role in helping citizens to comply, rather than merely conform, with moral reasons.

One of the most interesting features of PCA, and in my view the key to its potential in bringing about active or voluntary behaviour change, is that it operates by altering the way in which people see their choice environments. The main issue that interests us here is whether it is able to do this in ways that support the development and exercise of morally responsible agency. My argument in this section, and the central contribution of the chapter, is that there are three distinct ways in which PCA is able to encourage the kinds of moral improvement involved in exercising such agency. It is

[125] I thank Tom Parr for helping me to delineate these three defeasibility conditions more clearly.

argued that moral nudges can support morally responsible agency by: (i) helping people to act in line with their moral beliefs (as in (2) above); (ii) inducing moral deliberation (as in (3) above); or (iii) changing people's beliefs about what they owe to others (as in (4) above).

The first of these represents a form of motivational enhancement, since citizens already hold the relevant moral beliefs in these cases. For this reason, I call this category *motivational scaffolding* (§5.4.1). The second and third, by contrast, are primarily directed at improved insight – though they differ in terms of their (epistemic) directiveness. *Inducing moral deliberation* (§5.4.2) is relatively nondirective (which is why it can also act as a form of motivational improvement, as in (3a), as well as improved insight, as in (3b)), while *public-ecological persuasion* (§5.4.3) offers a more directive attempt to influence citizens' beliefs. Before discussing each of these in detail below, it will be instructive to offer some general comments.

Which kinds of moral duties might these interventions be directed towards securing? As mentioned above, the distinctive advantage and contribution of moral choice architecture rests in its ability to serve as a means of overcoming problems associated with *nonenforceability*.[126] There are many moral duties which are, for a variety of reasons, nonenforceable. We drew a distinction between two categories of such duties above: (i) those that are nonenforceable in principle because, by their very nature, they need to be performed freely; and (ii) those that are enforceable in principle, but which it is neither possible nor morally desirable for the state to coerce people into performing, perhaps because they are considered "imperfect duties" (Kant 2002: 40), or because the agent's decisions about whether or not to act on this duty are taken to be protected by an "injunction" (Kagan 1989: 219-230; more on this below). In both cases, there are good reasons – moral and ethical – to see that citizens fulfil these duties, even though it is not possible to use the traditional coercive measures of the state to do so. Our focus will tend to be on duties that fall into the latter category, those that are nonenforceable in practice rather than in principle, since this includes more of the kinds of duties on

[126] There are active, ongoing debates about how we should categorise many duties. One illustrative example is the debate over organ donation. We might think that there is a duty to make our organs available to those in need of life-saving transplants, at least in the event of our death. But many people think that this duty to aid is nonenforceable, that is, that the compulsory confiscation of cadaveric organs is wrong. Others have argued that the needs of the sick should, at least in principle, be met by the confiscation of organs from the dead and even from the living (Fabre 2006). There is, therefore, a certain level of *indeterminacy* about what types of duties some duties are (Krishnamurthy 2015). We will not enter into, or take any substantive views on, these debates here. Instead, we accept the less controversial view that we have many duties (e.g., duties to aid, duties of beneficence, and so on) that are putatively nonenforceable as the basis for exploring the conditions under which it might be permissible to use moral nudges as a means of compliance-promoting moral improvement.

which governments might feel that they need to encourage for the sake of meeting their public policy objectives: e.g., organ and blood donation, charitable giving for poverty alleviation, neutralising discriminatory implicit biases, pro-environmental behaviour to reduce the speed and harms of climate change, parents supporting children's learning, and so on.[127]

A separate distinction relating to moral compliance tracks, to some extent at least, the distinction between categories of nonenforceable duties. When a duty is non-enforceable in principle, it requires that we recognise and act from a particular (kind of) reason or motivation. In these kinds of cases, moral compliance is regarded as a duty, i.e., there are moral reasons to secure compliance. By contrast, in other cases, such as when a duty is nonenforceable in practice, it does not require compliance with reasons, though it remains in the interest of the agent to comply. In these types of case, moral compliance is not regarded as a duty, but as something that is valuable for the agent – she lives better by doing so. Given the focus noted above, it will be this second category – based on the ethical reasons (in addition to practical ones, given that coercion is not an option) in favour of pursuing moral improvement in a way that supports morally responsible agency – that will be our primary site of interest when considering how moral nudges might be compliance-supporting.

Such agency requires that an individual is both reasonably *receptive* (the epistemic condition) and adequately *reactive* (the control condition) to the moral reasons that there are.[128] To be morally responsible for an action, and thereby for it to have ethical value for the agent or for it to count as a fulfilment of compliance-requiring duties, the mechanism that issues in the desired or required action needs to be, in an appropriate sense, the agent's own mechanism (Fischer 2012: 187). How might choice architecture play a role in the design of policies, or "moral ecologies" (Kelly and Morar 2016), that meet these conditions for moral responsibility? One of the most promising approaches comes from taking into account the important psychological findings about the role of *salience* in attentional selection, and exploring the political implications of this empirically grounded view of moral psychology.

Within the study of perception and cognition, salience (or "saliency") is a term that refers to any aspect of an environment that stands out from the rest in some way.

---

[127] By comparison, the in-principle nonenforceable duties, which include the duties of friendship and consideration, are more controversial sites of interference since they relate more to private interactions and personal virtues.

[128] Here I follow Fischer and Ravizza in adopting a moderate version of reasons-responsiveness. For more on this, see Fischer and Ravizza (1998: 62-91).

As a result, salience – which can be the result of emotional, motivational, or cognitive factors – has a determining role in directing attention. For our purposes, the central thought is that there may be ways in which the political community could induce or support practical moral reasoning and reasons-recognition by making relevant moral considerations and reasons for action (more) salient within the choice environment. By designing choice environments that alter the way in which citizens see (perceive, feel, and so on) a situation, it may be possible to improve their understanding of the morally relevant reasons that apply to them in that situation. In some contexts this might be achieved by obscuring morally distracting or irrelevant features of the situation, so as to "hinder hindrances" to perceiving or acting in light of its morally relevant features (Connelly 2014: 228). The first of the three compliance-supporting moral nudges, which we move on to describing now, includes aspects of this hindrance-hindering, while the other two forms seek rather to induce moral deliberation and reasons-recognition.

§5.4.1: "Motivational Scaffolding"

Some nudges may support a person's ability or reduce her disability to act in accordance with how she believes she ought to act. In so doing, they can help to bring about greater coherence between an individual's behaviour and the belief she herself holds about her reasons for action in a given, morally relevant situation. In these cases, the person holds a particular belief, but, for one of many reasons including some relating to automaticity, she (predictably) fails to act in the way that she believes she ought. This mismatch between a person's beliefs and her behaviour can occur for a number of reasons. One is due to weakness of will (or moral *akrasia*), which captures cases in which a person holds a belief but experiences motivational hindrances to acting on it. Another relates to lack of awareness: although an individual holds the belief that she should act in a certain way in relevant circumstances, she might be unaware that the current situation is such a circumstance – something that is often due to attentional deficits. A third is on account of the phenomenon of (inauthentic) implicit bias, which issues in automatic prejudiced judgments and social behaviour, even in those who explicitly hold non-discriminatory beliefs.

It is clear that the harm and wrongs committed would be significantly reduced if people could more reliably (be motivated to) comply with the moral reasons that they acknowledge and accept. But, in addition, part of the justification for motivational scaffolding comes from its potential to promote the agent's morally responsible agency in so far as it promotes her autonomous agency. The three examples above describe

instances of there being a mismatch between an individual's beliefs and her actions, and such mismatches are generally considered to be a problem for her autonomy. Any motivational improvement that PCA can generate would serve the individual's interest in being able to execute the moral beliefs she has. This reasoning grounds a (defeasible) reason in favour of these compliance-supporting moral nudges.

This argument, however, requires a content-sensitive limitation because the ideal of autonomy, and its associated weight as a moral constraint to intervention, does not extend from the prudential to the moral domain in any straightforward way. In short, it matters whether an individual's beliefs reasonably track the moral reasons there are. Even though it would promote a person's autonomy in a particular sense if she is better motivated to act on her moral beliefs, if these beliefs are inadmissible then interventions that improve her motivation to act on such beliefs do not support moral responsibility, and so do not contribute to living well at all. Another way of putting this is: if (and only if) a person is *receptive* – even if not adequately reactive – to a decisive moral reason for action, then nudges that help to reduce the mismatch between her belief and her action can support moral responsibility and, consequently, can play a role in supporting her in living well.

Unlike the two categories of moral nudges that are directed at improved moral insight (detailed below), motivational scaffolding could, in principle, take the form of either a "System 1" or "System 2" nudge, although additional constraints may be necessary in the case of the former. Let's begin with the case of neutralising implicit biases, where System 1 nudges are most likely to be permissible on account of the nature of the problem at hand. Implicit biases are "automatic associations, often operational without the reflective awareness of the agent, which influence action" (Holroyd and Kelly 2016: 106). As Moles highlights, the "unconscious nature of implicit biases does not cancel the duty that individuals have not to behave in a racist [or otherwise wrongfully discriminatory] manner" (2015: 663). The most potent aspect of these biases is that they are possessed by most people, including those who explicitly and sincerely hold egalitarian beliefs – a phenomenon that has been called "aversive racism" in cases where the bias is based on racial associations (Pearson et al. 2009). The issue of whether and how we might be able to regulate and control these biases is currently a site of much empirical and philosophical interest (see, e.g., many of the chapters in Brownstein and Saul 2016a and 2016b). What the empirical research does suggest is: (i) that attempts at directly suppressing implicit associations are unlikely to be

successful and, in fact, may sometimes exacerbate the problem (Holroyd 2012 on the evidence on "rebound effect"), which seems to advise against System 2 nudges; and (ii) that individuals' ability to sustain and implement their egalitarian commitments in the face of these unconscious threats "depends crucially on hospitable environments" in helping to *automatise* the egalitarian motivations that are representative of her moral beliefs (Rees 2016: 209-210).

Taking these two things together, motivational scaffolding is likely to work best when delivered via System 1 nudges in the case of reducing implicit biases. For instance, the political community might arrange social environments in ways that offer counter-stereotypes, the idea being to rewrite citizens' automatic learned associations over time in order to dissociate the negative link between, e.g., black people and violence (Dasgupta 2013; Moles 2015: 664).[129] There is also some evidence about regulating the cognitive "accessibility" of stereotypes: for those with egalitarian beliefs, there is some evidence that biased reactions can be partly regulated by nonconsciously "activating" these goals (Madva 2016).[130] This is because, given limited cognitive resources, people struggle to keep their liberal and egalitarian goals activated all the time, with the result that stereotypes are more readily accessible in the brain. Hence, PCA may have some role to play in "scaffolding" moral cognition by keeping egalitarian goals activated.[131]

Despite sharing a similar underlying motivation, these System 1 nudges differ in subtle but important ways from the auto-nudges discussed in the analysis of prudential nudging. That category was restricted specifically to System 1 nudges that a person herself opts in or consents to. As we have been focused on that category of people who are committed to non-discrimination above some threshold, it is likely that these people would consent to regulating their implicit biases in these ways. But, in the moral realm (or in cases like this one at least), the commitment to consent is lessened, because these interventions are designed primarily for the sake of preventing harm to those who are the target of these implicit biases. Those who do not hold egalitarian beliefs therefore

---

[129] The evidence shows that white people, even those with explicit egalitarian beliefs, "often display reflexive signs of distrust and aversion, including increased blinking rates and decreased eye contact when they interact with black people" (Hueber 2016: 48; for original study, see Dovidio et al. 1997). This leads to epistemic injustices, in particular, testimonial injustices (Fricker 2007) and can have serious – even life-threatening – implications for black people, such as the well-documented "shooter bias" (see Brownstein 2016: 776-778 for an overview of the evidence).

[130] In psychology, *accessibility* refers to "the ease with which a particular unit of information is activated or can be retrieved from memory" (Morewedge and Kahneman 2010: 435; cited in Madva 2016: 192).

[131] The closest Sunstein gets to talking about moral nudges in the way that we are understanding them is a short clarification in his latest book, *The Ethics of Influence*, in which he states that: "Some nudges are designed not to promote the interest of choosers, but to give fairness the benefit of the doubt, by promoting particular understandings of (say) equality" (2016a: 34-35).

do not have a complaint against their use. Nonetheless, the hypothetical consent of the portion of the population for whom these interventions count as a form of motivational scaffolding, a significant proportion if the existing data is anything to go by, is not unimportant; it signifies that moral nudges of this kind are compliance-supporting for these citizens. These nudges deliver on Sunstein's promise that nudges increase how well a person's life is going by "steering [his or her] decisions in the right direction (*as judged by themselves*)" (2014a: 17; emphasis added).

There are examples of motivational scaffolding that do operate in a way that parallels the auto-nudge mechanism. One example is the weak-willed philanthropist (Krishnamurthy 2015). Consider a person who holds the belief that the needs and human rights of those in poverty mean that she ought to give a certain amount of her disposable income each month to effective poverty-alleviation charities. Even though she sincerely holds this belief, she is not always adequately motivated to make the transfer that she believes she should make. It is permissible to do a number of things that might help this individual act on her beliefs by helping her to overcome her weakness of will. The charity to which she donates, for instance, could give her the option of setting up a monthly direct debit transfer or opting in to a voluntary text message reminder service that sends her a message at the beginning of each month reminding her about her charitable commitment – these are classic examples of pre-commitment or "auto-nudges". Similar kinds of policies could be used for supporting a person's commitment to, e.g., donating blood at regular intervals, carbon offsetting, and so on.

In one way, such policies would seem not to be compliance-supporting. If people are weak-willed or attentionally deficient in certain situations, and need PCA to help them secure certain outcomes in practice, it would appear that they are not sufficiently reactive to reasons, even if they do recognise them. But this concern misses the point of these kinds of interventions: in the empirical conditions under which we seek to exercise our moral agency, recognising that there are several kinds of motivational threats to bringing our moral beliefs into action and acting in order to lessen or cancel the effects that these threats might exert on us can be seen as an important part of what it means to exercise morally responsible agency. Callan expresses a similar point when he states that, "A necessary part of virtuous motivation as it has been traditionally conceived is the desire to avoid circumstances in which our own virtue is needlessly put under strain" (2015: 493). PCA offers one way in which we can

support this reasons-reactivity side of moral compliance, which links back to our earlier discussion (in §4.4.1) about the idea that autonomous and now, by extension, morally responsible agency is partly a social achievement, as well as an individual one.

So far we have discussed System 1-based versions of motivational scaffolding. This relates to the second potential aim of moral nudging, i.e., (2). It is also possible that System 2 nudges can act as a form of motivational scaffolding. One of the potential outcomes of inducing moral deliberation – the category explored in full in the next subsection – is that it might help people to bring their own behaviour in line with their beliefs in morally relevant situations, i.e., (3a). By encouraging people to direct their attention towards some morally salient feature of a choice environment, choice architects might not only nudge citizens into taking into consideration reasons they had not previously recognised, they might also remind those who have already recognised the relevant reasons of their commitment to act in accordance with them. Of course, choice architects cannot discriminate between these two groups; but there is no need to, since the same intervention will interact with people's practical reasoning processes in different ways according to their existing epistemic status vis-à-vis the issue at hand.

One limitation of motivational scaffolding is that it is premised on citizens' already holding the relevant moral beliefs. This shifts our attention to the question of how moral nudges might act on citizens who do not. The second and third categories in the account are concerned with PCA's potential as a means of enhancing an individual's *understanding* of morality. This is more controversial than motivational scaffolding because, rather than seeking to bring a person's behaviour in line with the beliefs she already holds, these interventions seek to influence and develop her understanding of the moral reasons there are and, in some cases, to do so in a way that seeks to change her current beliefs about what she has reason to do. There are two types of roles that moral System 2 nudges might play in this process: inducing moral deliberation (discussed in the next subsection) and a more directive version that I am calling "public-ecological persuasion" (discussed in §5.4.3).

§5.4.2: "Inducing Moral Deliberation"

Like the nudge-autonomy mechanism outlined in Chapter 4, moral System 2 nudges can support compliance by encouraging, prompting, or inducing moral deliberation on a particular issue about which citizens are reasonably considered to have a duty. One of the main ways of doing this is to make its morally relevant aspects more salient within

the choice environment.[132] Consider the policy for organ donation registration in the UK. Unlike other countries that have either an opt-in (e.g., Germany) or opt-out (e.g., Austria) default rule, the UK has designed an active choosing policy that asks people when they visit the DVLA website to apply for a driving licence or to tax their vehicle whether they wish to register as an organ donor and, if so, which organs they would be willing to donate in the event of their death. This type of design encourages people – specifically, those who are not already registered as donors – to think about the issue. We might assume that, for some at least, this encourages them to deliberate to some degree about what it is that they owe to others and how this should be weighed against other, prudential reasons they might have.

This type of moral nudge is nondirective in an important sense. Although it is a measure that is intended to increase registration, and thereby to lessen the harm suffered by those in need of transplants, it does not seek actively either to manipulate or persuade people into making a particular choice. Rather, it respects that people are capable of reasons-responsiveness and that, either via individual forms of deliberation or by deliberating with others about what it is reasonable for citizens to expect of one another, citizens can come to appreciate the applicability and force of relevant moral reasons. The deliberation that is elicited may clarify, and in some instances change, an individual's belief about what he or she owes to others in this particular case. Sometimes this will involve the application of a belief relating to a moral principle – such as a general commitment to fairness and reciprocity, for instance – to a specific real-world case, thereby producing a more specified belief about what such a commitment means for one's decision about whether or not to register as an organ donor. At other times, it may lead to citizens forming or modifying a belief at a more abstract level by creating an opportunity for people to "grapple with the relevant moral concepts directly" (Shiffrin 2010: 1222). In either case, these nudges can support moral compliance and, as a result, can serve the interests of the agent, as well as the interests of those (who will be) in need of organ transplants.

---

[132] Note the difference between aiming at encouraging moral deliberation and aiming at another good, but foreseeing that the most effective choice architecture design might act by nudging people into moral deliberation. This relates to different uses of the term *intention* (and therefore *intentional*): in both cases we are aware of what will be the consequences of the intervention (the broader sense); but only in the first case do we intend moral deliberation in so far as this is what we aimed at or planned for (the narrower sense) (Scanlon 2010: 10-11; see Anscombe 1958: 9, for a similar distinction). This is related to the point made in fn. 78. Here, we adopt the broader notion (which includes the narrower sense), because the principal goal is to protect others from harm, even though there may be ethical value to doing this in a way that encourages people to recognise and respond to moral reasons.

In reality, the deliberative processes that result from these nudges will only rarely, if ever, approximate the idealised – or rather "unrealistic" – model of deliberation that is thought to involve an individual identifying an evaluative term, coming to a considered judgement about its proper application, and then adjusting his or her conduct accordingly at that point in time (Shiffrin 2010: 1226). But the fact that the induced deliberation will include unconscious as well as conscious elements (such as emotions) and errors as well as insights, both across the citizenry and within individual citizens, does not offer a reason against the use of nudges that aim at influencing people's beliefs and behaviour in this way.

Inducing moral deliberation might have an interesting role in counteracting a phenomenon known as "strategic ignorance" (Wieland 2016, 2017; Heffernan 2012 explores a similar phenomenon, calling it "wilful blindness").[133] The central thought here is that sometimes we know (in a second-order sense, as it were) that being aware of certain facts about a situation (in a first-order sense) would give us reasons to change our current pattern of behaviour; thus, we deliberately try to avoid such knowledge, most likely because we would feel forced to weigh the moral reason against conflicting, self-interested ones. This example is difficult to categorise because in some ways it is similar to cases requiring motivational scaffolding: an individual has a particular moral belief or is aware of a moral belief that she thinks she *should* hold, and the aim of the nudge intervention is to try to reduce the mismatch between this and her behaviour. But it also seems that there is something different about these kinds of cases. In particular, there is a sense is which the nudges that act as motivational scaffolds in the cases above secure or at least support moral compliance in a way that they would not do in cases of strategic ignorance. The difference seems to come from the strategic nature and perhaps even intentionality associated with the ignorance, namely, this betrays a lack of commitment to the moral belief and the role that it should play in our action. There is a certain amount of moral insight in these cases, but this insight is being used more to shield the person from some inconvenience or their anticipated guilt rather than to direct their actions in morally good ways. Strategic ignorance does not cancel out the duty that individuals have not to be complicit in, for example, exploitative and unjust labour conditions; so nudges that make certain morally relevant information

[133] Moody-Adams has suggested that "the main obstacle to moral progress in social practices is the tendency to widespread affected ignorance of what can and should already be known" (2004: 266). Wieland offers an analysis of "willful" (affected, motivated, strategic) ignorance (2017), and discusses issues relating to responsibility for such ignorance (2016).

salient can induce (unwanted) moral deliberation about how the moral reasons should weigh against other reasons.

To clarify, the principal aim of using these kinds of nudges is, like those belonging to the nudge-autonomy mechanism, to help citizens better to recognise *for themselves* the moral reasons that apply to certain situations. They do not seek to direct or persuade citizens towards particular decisional outcomes, despite there clearly being a sense in which the choice architecture communicates something about the political community's norms and judgements. Shiffrin makes a similar argument with respect to legal standards (vis-à-vis legal rules), which I made reference to in §5.1. She writes:

> "the rationale for standards I have been discussing does not aim to impart specific content through bypassing independent deliberation but rather aims to spark independent deliberation; although the standards themselves incorporate moral concepts and so, in that way, aim to elicit an interpretation or conception of what is morally apt or appropriate, that standard that elicits deliberation dictates no specific content to that interpretation." (Shiffrin 2010: 134)

In the organ donation example, for instance, the main aim is that people are responsive to their own reasoning on this issue, which is an important feature of morally responsible agency too. As it is, policies of this kind are directive in so far as they direct people's attention and deliberation toward an issue and in so far as this is perceived by citizens as carrying an implicit recommendation about behaving or deciding in one way or another. But people are free to make their decision about whether to register as an organ donor without any further pressure being exerted on them. There is another category of System 2 nudge that does seek to exert this kind of persuasive pressure for the sake of enhancing citizens' understanding of morality. We discuss this next.

§5.4.3: "Public-Ecological Persuasion"

A more controversial type of moral nudge might seek not to bring behaviour into conformity with one's beliefs or to encourage moral deliberation, but to change people's beliefs concerning what they owe to others. Its controversiality rests on the idea that interventions that interfere with people's belief-formation are often more worrisome than those that influence their actions (Clayton and Moles, forthcoming). Recall that there is an epistemic condition, as well as a control condition, for morally responsible agency. The issue at stake relates to whether it is permissible for the political community to influence an individual in ways that enhance her understanding of morality for her

own sake, that is, in ways that help her to meet the epistemic condition required for moral responsibility. The challenge that needs to be addressed is: even if moral responsibility is part of what it means to live well, and that this requires moral insight, it does not necessarily follow that it is permissible for the state to seek to morally educate its (adult) citizens without their consent. In response, I will defend the view that moral choice architecture can act as a form of persuasion – a form that I will call *public-ecological persuasion* – and that its use as a means of belief and behaviour modification is morally permissible under certain conditions.

There is surprisingly little philosophical work on persuasion as a political method.[134] This dearth is odd for two main reasons. First, despite rational persuasion generally having a special moral status as compared to other means of influencing beliefs and behaviour, there is not much more than a common characterisation of it in the political-philosophical literature as, in short, *reason-giving*.[135] Second, there have been substantial developments in both psychology (e.g., Cialdini 1994) and technology (e.g., Fogg 2003) that have provided us with the tools both to understand and to exploit, respectively, the power of influences on our beliefs and behaviours. These have often been described in terms of persuasion – e.g., "persuasive technologies" and Packard's (2007) description of advertisers as "hidden persuaders" – despite their questionable connection to reason-giving. One way of describing our discussion here is that it will take place within the conceptual space created by the interaction of these two reasons: based on the psychological evidence about influences on our beliefs and behaviour, it will explore the potential for a special type of rational persuasion that extends the common characterisation of reason-giving to include reasons that are given by being built, in effect, into the choice architecture. It will argue that this type of persuasion, under certain conditions, shares in the special moral status that is given to the other members of this family – even if only in a more qualified way.

Rational persuasion involves "the act or activity of presenting information, facts, evidence, or arguments that one [i.e., the persuader] takes to provide reasons", where reasons are "particular considerations that count in favour of belief, attitude, or action, and that have normative force" (Tsai 2014: 89). Thus, the activity of offering reasons to a person represents an attempt on the behalf of the persuader to change the receiving

[134] Some notable exceptions include Burnell and Reeve (1984), Strauss (1991), Sawicki (2016), and Tsai (2014).
[135] For a recent paper questioning this special moral status, see Tsai (2014), which argues that some kinds of rational persuasion are paternalistic.

agent's beliefs and his or her subsequent behaviour.[136] Through the provision of reasons, persuasion aims to influence how people see their choices in relevant choice environments. The reasons given can be either *explicit*, as when a doctor tells an individual that her (deteriorating) health gives her reasons not to smoke, or *implicit*, as when a doctor presents her with statistics relating to the relationship between poor health and smoking.[137] It does not appear that there is any normative difference between implicit or explicit forms of rational persuasion.

In their assessment and critique of nudging, Hausman and Welch claim that: "Coercion is often justified, and shaping [i.e., PCA] sometimes a better alternative than coercion, but rational persuasion is the ideal way to influence the behaviour of citizens" (2010: 135). However, is the distinction between rational persuasion and nudges (i.e., "shaping") as stark as these commentators make out? Their claim about the relative desirability of the different interventions rests on the view that "only rational persuasion fully respects the sovereignty of the individual over his or her own choices", that is, his or her autonomy (ibid.). Hausman and Welch nonetheless note two important qualifiers concerning (i) the limited force and effectiveness of rational persuasion and (ii) the fact that, in reality, persuasion is "rarely purely rational". The authors follow the second up with a footnote, important within the context of the view I wish to defend, which states that:

> "A systematic account of rational persuasion is needed here. We do not mean to suggest that rational persuasion is emotionless cold calculation. *Clarifying the role of emotion in rational persuasion* is a difficult task for another occasion." (Hausman and Welch 2010: 135; emphasis added)

It seems clear, to me at least, that the concept of persuasion is in need of further conceptual refinement, especially in light of the psychological evidence and its bearing on how we conceive of rationality; and that the corollary of this is that its use as a

---

[136] The use of the term "attempt" is important. It refers to a distinction, which Burnell and Reeve highlight (1984: 399), between two ways in which we might understand persuasion. On the *success* conception, calling something an act of persuasion is dependent upon it producing a successful outcome, namely, that the target comes to share the reasons advanced by the persuader. See, for instance, Faden and Beauchamp's approach which refers to persuasion as "an intentional and successful attempt" (1986: 261, 347). By contrast, the *process* conception views persuasion simply as the activity of advancing reasons, regardless of whether the target is persuaded by the reasons or not. This distinction is important because it determines persuasion's relative distance from coercion. One way of thinking about the nature of coercion is that it simply fails to be coercion if it is unsuccessful; that is, "coercion depends upon success" (Burnell and Reeve 1984: 396). The success conception therefore makes persuasion more like other forms of power, such as coercion, than perhaps we might think it should be. For this reason, I adopt a process conception.

[137] I first saw this distinction, together with a similar example, in an unpublished paper by Tom Parr.

political method is also in need of further normative analysis. Although this "difficult task" cannot be addressed here in the detail it deserves, the argument offered in this subsection is intended as an initial contribution to developing this part of the literature, specifically as it relates to the ethics of behaviourally-informed governmental influence.

We first need to address a terminological issue. Some use "persuasion" in a broad sense that takes it away somewhat from the central idea of reason-giving. In a recent paper that marks one of the few normative contributions to the literature, Sawicki distinguishes between different kinds of "persuasive appeals" according to the "degree to which they engage reason". She maintains that some are made on rational grounds (such as "presenting truthful facts […] with no inaccuracies, omissions, or biases"), while others influence choices "on grounds unrelated to reason – because they distract from rational deliberation, operate on a plane apart from rational deliberation, or render rational deliberation irrelevant" (2016: 213). On her definition, persuasion that is based on "arational" factors "either distracts from, or renders irrelevant, the listener's reliance on reason in making a given decision", and she lists emotional appeals and appeals that rely on cognitive biases alongside subliminal messaging as examples of this arational persuasion (Sawicki 2016: 214).

This raises a number of issues that I wish to address, in particular, the role that emotions can play in our reasoning processes. The point to make at this stage, though, is that what Sawicki is defining as arational persuasion is more appropriately described in terms of "manipulation", given the way in which she has described what it means for an influence to be arational (i.e., as reason-bypassing in various ways). Accepting this allows us to keep the term "persuasion" for reason-giving influences. This is the terminological route taken by Tsai, too, whose work also highlights that manipulative influence can take place via rational means. He gives the following examples of *rational manipulation*:

> "One might introduce reasons into a person's deliberations in order to play on her neuroses (or simple, individual preoccupations) so as to impede her deliberation and control the likelihood of a certain outcome to her deliberation. One might inundate someone with lots of relevant information, doing so precisely in order to overwhelm her and hinder her deliberation. One might present someone with evidence with the aim of emphasizing certain relevant information and facts, while intentionally neglecting to mention other information one acknowledges as relevant." (Tsai 2014: 89-90).

As a result, we can represent this conceptual space with four quadrants, where the *x*-axis captures the space between manipulation to persuasion, and the *y*-axis captures the space between rational to nonrational (not "arational") forms of influence (see Figure 1, below). Four kinds of influence are roughly plotted on this graph: those that tend to be viewed as classic cases of rational persuasion (e.g., factual information provision) and of manipulation (e.g., subliminal messaging) are marked by X_1 and X_2, respectively; and those that count as less well established forms of both persuasion (i.e., public-ecological persuasion) and manipulation (i.e., rational manipulation) are marked as X_3 and X_4, respectively.

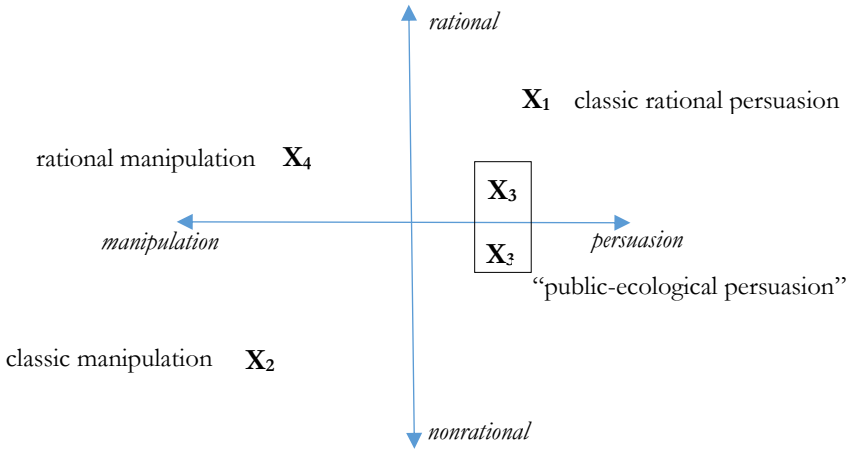

Figure 1: Plotting the conceptual space.

The first point to make about this conceptual space requires returning to Sawicki's conceptualisation. Recall that she defines "arational" persuasion as those influences that affect choices in ways that are "unrelated to reason – because they *distract* from rational deliberation, *operate on a plane apart* from rational deliberation, or *render* rational deliberation *irrelevant*". We have categorised such influences instead as manipulation. Where Sawicki's conceptualisation goes wrong, in my view, is that it appears at times to categorise all forms of influence with emotional and cognitive heuristic-based elements as arational and, consequently, as manipulative.[138] But there is an important difference between influences that act in this *arational* or manipulative

[138] Sawicki claims, for instance, that: "While subliminal advertising may be an extreme example, the same outcome would likely play out were the state to communicate through appeals to emotions or cognitive biases – a person's thought process would be affected without her consent, and she would not be able to make a substantive decision without the intrusion of the government's message" (2016: 218-219).

fashion, and those that involve *nonrational* elements. On my view, there are ways in which influences that include cognitive heuristic-based and emotional elements can permissibly be used with the aim of engaging reason, for the purpose of helping people better to appreciate moral reasons. This is the basis of the nudge-based category of persuasion that I label *public-ecological persuasion*.

Public-ecological persuasion will often take the form of what we might call "enhanced" active choosing policy designs.[139] Take as an example a prompted choice design for organ donation registration that includes a factual statement intended to make salient the moral reason in favour of registering as a donor. One possible example would be: "On average, 22 people die each day in the United States alone while waiting for a transplant".[140] Another might make use of a descriptive norm, such as stating the percentage of people who are already registered as organ donors. There is clearly an element of rational persuasion here; indeed, some would be happy to call such an influence just that. My own view is not very far away from this, as can be seen in where this influence is plotted on the graph above. But it does contend that public-ecological persuasion is a distinct category. In particular, these persuasive efforts are better described as being a kind of mixed policy: part rational persuasion, part nudge. They clearly involve efforts at rational persuasion because they convey important information that is relevant to the person's decision, and they allow an individual to engage in their own deliberative process about how to respond to this information. Nonetheless, designing an active choice framework (in place of, say, the traditional opt-in framework) and presenting the information in psychologically salient ways – such as by framing the information in terms of those who die waiting for organs rather than those who successfully receive organ transplants, or in terms of those who are already registered as donors rather than those who are not (Bovens 2009: 208-209) – clearly also count as elements of nudging, that is, as using public-ecological design as a means of bringing about belief or behaviour change.

Some might worry about this "persuasive design" of moral nudge policies on one of two potential grounds. The first concerns their directiveness. In the case of prudential nudging, it was deemed objectionably paternalistic to use System 2 nudges to steer people's deliberative processes in a directive way. Why, then, is this not the same

[139] Sunstein and Reisch briefly introduce (a version of) "enhanced" or "influenced" active choosing as a possibility (2015: 154), though this idea is originally introduced by Keller et al. (2011).
[140] See the following webpages for infographics displaying this figure from the US Department of Health and Human Services (2016): (1) https://www.organdonor.gov/statistics-stories/statistics.html, and (2) https://optn.transplant.hrsa.gov.

in the case of moral nudging? The answer can be found in the differences between the epistemic conditions for autonomous and morally responsible agency. As was explained in §5.2.2, supporting the latter requires supporting citizens in coming to an appreciation of the moral reasons, but doing so in a way that gives them the freedom to see things aright. The aim of the System 2 nudges that are used in public-ecological persuasion is precisely this: to encourage citizens to recognise and appreciate moral reasons, where fully appreciating a reason means that an agent assigns it the appropriate weight in her deliberations (Hanna 2015: 627-628). The asymmetry with the prudential realm is grounded in the following three points: (i) the state does not have the same authority to protect people from their own self-regarding behaviour (when this is sufficiently voluntary and autonomous) as it does in constraining the harm that their other-regarding behaviour can cause to others; (ii) this is partly because there is and always will be considerable reasonable disagreement about what it could mean to "see things aright" in the prudential realm; and (iii) this heterogeneity is central to the idea that it is part of what it means to live well to take an active role in creating and performing one's own life, which is what grounds the value of autonomy. (I return to this concern in the next subsection.)

The second kind of concern worries not about the directiveness itself, but about the mechanism by which it operates. Does it work by inducing certain emotional reactions? And, if so, does this count as a form of psychological manipulation? The worry is that the addition of the statement about the average amount of daily deaths of people waiting for organ transplants is intended to work, and indeed is successful at generating behaviour change, by inducing guilt in citizens about not registering as organ donors. If this is true, then it may seem that it is not their appreciation of the reason that produces the outcome, but rather the fact that the government intentionally played on their emotions in a potentially manipulative way in order to bring about a particular decision. I think this concern is an important one; if this were the case, this would raise interesting questions about whether this might still be permissible. Nevertheless, I also think that this mischaracterises the way in which affect-cum-motivational states such as guilt are involved in bringing about behavioural changes in cases such as these, and that clarifying this diffuses the worry (and, more than this, offers further support for public-ecological persuasion).[141] To show this, I will engage with some recent studies that seek

[141] For an instructive collection of essays on the relationship between morality and the emotions, see Bagnoli (2011).

to explore the underlying mechanism at work in nudge policies aimed at reducing harm to the environment.

The early work in this area has tested experimentally to what extent it is guilt – or, more precisely, "anticipated guilt" – that is responsible for bringing about pro-environmental behaviour change (Theotokis and Manganari 2015; Hedlin and Sunstein 2016). One of these studies tests how active choosing policies (some "enhanced", others not) fare in terms of their effectiveness at getting people to choose the "green" over the "grey" energy program, as compared with automatic enrolment in the more climate-friendly option (Hedlin and Sunstein 2016). The authors report three main findings: (i) active choosing led to higher enrolment in pro-environmental behaviour; (ii) active choosing caused participants to feel more guilty about not enrolling in the green energy program; and (iii) the level of guilt was positively related to the probability of enrolling. How might we interpret these findings and, in particular, the role that guilt plays? We might distinguish between three interpretations. The first, which is the basis of the concern outlined above, is that guilt is being used as a policy tool to bring about behaviour change via the motivation of guilt-avoidance. On this interpretation, the difference between active choosing policies and automatic enrolment defaults is that the former, in prompting deliberation, seeks intentionally to induce or trigger feelings of guilt about not acting in a pro-environmental way. Since guilt is "an aroused form of emotional distress that is […] based on the possibility that one may be in the wrong or that others may have such a perception" (Baumeister et al. 1994: 245), this also induces a desire in people to get rid of this unpleasant feeling and, in so doing, motivates them to conform to acting in the desired way.[142]

It seems to me that this is not a particularly plausible interpretation of the results. This is due mainly to my account of moral psychology, which includes the perceptual theory of emotions that I introduced earlier in the chapter. According to this theory, emotions can play a role in informing us about our moral reasons. The least controversial type of claim about this epistemic function, supported by a significant amount of empirical research (Damasio 1994), is that emotions act as detectors of salience.[143] As De Sousa explains:

---

[142] This type of explanation is referred to in Theotokis and Manganari's article, where they state that "consumers are stimulated to act against the cause of the guilt, in order to alleviate the harm caused and consequently the feeling of guilt" (2015: 425).

[143] Experts in this area state: "We consider affective processing to be an evolutionary antecedent to more complex forms of information processing; but higher cognition requires the guidance provided by affective processing" (Adolphs and Damasio 2001: 45).

> "Despite common prejudice, reason and emotion are not natural antagonists […] For emotions are among the mechanisms that control the crucial factor of *salience* among what would otherwise be an unmanageable plethora of objects of attention, interpretations, and strategies of inference and conduct." (De Sousa 1997: xv; emphasis is original)

Contra those who regard emotion as "at best irrelevant to, and at worse as distorting, good reasoning", this empirically grounded view suggests that "affective responses are actually, though only partially, constitutive of good reasoning" (Levy 2017: 5; see Berker 2009 for more detailed discussion).[144] The central idea is that emotions, in this case the feeling of anticipated guilt, draw our attention to morally relevant features of the situation, which enables deliberation and, as a result, can help people better to recognise and appreciate reasons that apply to them.[145]

Hence, even if these kinds of nudges involved (reported feelings of) guilt and that this is important for explaining people's changes to their behaviour, this is not necessarily problematic in the way specified by the concern. Enhanced active choosing policies may "urge consumers to consider their ethical responsibilities" (Theotokis and Manganari 2015: 426), or "trigger moral values and social norms" that people can then deliberate on (Hedlin and Sunstein 2016: 115). This is permissible in the moral realm when it triggers emotions that serve to highlight reasons that the person accepts or could accept (and she accepts the salience of these reasons to the choice at hand); it can support citizens in exercising morally responsible agency. This would be enough to diffuse the concern, on the assumption that results show that guilt is causally involved in bringing about the behaviour change. Nonetheless, it is not clear that we need to accept this assumption, as we shall see below.

The third interpretation contends that the finding that active choosing is more effective at producing compliance than green defaults is due to the simple fact that, when people are required to choose, that mere requirement "might trigger otherwise

---

[144] This more subtle view accepts that emotions do not always lead to better deliberation and decisions. As Levy contends, "The naïve view is quite right that sometimes strong emotions may cause people to take unjustifiable risks, overlook important options, and so on. But objecting that the influence of affect on cognition cannot amount to reasoning on the grounds that it sometimes misleads us is holding it to a standard that *nothing* meets. Conscious reasoning, too, may mislead…" (Levy 2017: 5).

[145] Some have made stronger claims about the relationship between emotions and reasons-responsiveness, namely, that "acting on the basis of an emotion may involve responding to a reason" (Tappolet 2014: 171; Tappolet 2016; Jones 2003). The less controversial, "majority view" (which I am adopting) is that: "our emotions do not, as such, allow us to respond to reasons. When we undergo an emotion, we, or maybe more accurately some mechanism in us, might at best be tracking reasons, but we do not respond to reasons in virtue of our emotions. What is taken to be necessary for responding to reasons is deliberation (i.e., reflection about practical reasons), and this, it is thought, requires judgements about the agent's practical reasons" (Tappolet 2014: 171; for advocates of this type of view, see Korsgaard 1997; Wallace 1999).

dormant or ineffective moral values and social norms", and especially so when the active choosing design is "enhanced". When asked a question about what to do in such a situation people's response, according to Hedlin and Sunstein, tends to be in line with "what they believe morality requires" (2016: 137).[146] It is the thought of diverging from this in their decided-upon action that induces the reported feelings of guilt; but, importantly, this occurs only as a *by-product* of recognising the moral reasons that apply to them in this situation. Despite the reported feelings of anticipated guilt, then, this emotion does not play any causal role in changing the agent's behaviour. The aim of policy designs that work in this way is to make certain reasons that people might have to act in a particular way more salient within the choice environment. The reported feelings of anticipated guilt might be understood as evidence, therefore, that the policy is working with respect to making people consider and respond to the reasons to act in a pro-environmental way. In other words, guilt might have probative importance in so far as it is an indicator that active choosing policies improve compliance. If accepted, this interpretation nullifies the concern about guilt-inducement playing a causal role in behavioural public policy.

This illustrative example has highlighted two possible responses to this second concern. For my argument, it does not matter whether one finds the second or third interpretation more convincing. My sense is that, outside of this particular example, some forms of public-ecological persuasion will involve mechanisms that are explained better by the second interpretation, while others will operate in ways better captured by the third interpretation. The general point is that even if public-ecological persuasion involves nonrational elements (as when it works according to the second interpretation), this does not make it a form of "arational persuasion"; rather, it is a form of non-classic rational persuasion, where the PCA has been designed in ways that indirectly offer reasons by making them more salient within the choice environment.

You will notice on the graph above that public-ecological persuasion is plotted as sitting across the rational and nonrational boundary. This is because it will sometimes involve nonrational elements, though these are for the sake of engaging the individual's moral reasoning; and sometimes it will, more simply, engage her reasoning processes along the line of more classical rational persuasive appeals. In both cases, efforts at public-ecological persuasion are permissible in the moral domain, and offer a promise

[146] For clarity's sake, it is worth noting that Hedlin and Sunstein (2016) do not offer a clear or explicit position with respect to the underlying mechanisms at work; here I am reconstructing possible positions, some of which are hinted at by passages in their discussion of their results.

of helping governments deal with problems that come about as a result of the non-enforceability of some moral duties.

§5.4.4: Weighing Compliance and Conformity

Paralleling the argument made in the previous chapter concerning PCA designed to support citizens' prudential practical reasoning, this section has outlined ways in which moral choice architecture might be used to support citizens' practical moral reasoning. It has defended three kinds of moral nudge as presumptively permissible: motivational scaffolding, inducing moral deliberation, and public-ecological persuasion. The first is a System 1 nudge that supports autonomy by helping citizens to reduce the mismatch between their moral beliefs and behaviour. The second and third are kinds of moral System 2 nudges that aim at moral improvement principally via improved insight, in particular, by supporting people in being sensitive and responsive to the reasons provided by the value of other people's lives. Any behavioural improvements that follow as a result of this will represent both autonomous and morally responsible agency, meaning that they will have ethical value for the agent, as well as the moral value that comes with duty-fulfilment.

But this raises an important question, flagged earlier in the chapter, relating to cases in which there is a conflict between intervening for the sake of behavioural improvement (only) and intervening for the sake of improving an individual's ability to act in a morally responsible way. We have seen that there are reasons to favour nudges that support rather than bypass our practical moral reasoning. But can it ever be permissible to sacrifice some – or indeed, any – conformity with moral duties if doing so would be the inevitable by-product of individuals achieving more compliance? With regards to the category of nonenforceable duties that these moral nudges are aimed at, this weighting issue would arise in those cases where the choice architect has the option of choosing between different nudge designs in a way that takes on the following form: *Nudge A* (e.g., a System 1 nudge) improves behavioural outcomes more than *Nudge B*, but does so without improving her moral reasoning, whereas *Nudge B* (e.g., one of the three moral nudges outlined above) is better in the latter respect but simply less effective in achieving the morally desirable behavioural outcomes than *Nudge A*.

There are many different considerations that would need to be taken account of in assessing this type of issue, and it is not likely that there is any general answer that would apply in all such cases. It is important to make clear that my argument in favour of autonomy- or compliance-supporting nudges does not overreach. It acknowledges

that, although there are reasons to favour nudges that support rather than bypass our practical moral reasoning when all else is equal (when they are both as effective as each other), it is clear that all else is not often equal in real-world cases, given that System 1 nudges, being a type of enforcement (as we suggested in §5.3), are more likely to secure moral conformity than System 2 nudges. This will not always be the case. In some cases, measures that offer the opportunity for moral compliance may be *more* effective at bringing about the desired behavioural change than those that do not. We saw this in the data presented above on pro-environmental behaviour, where active choosing policies secured the desired behaviour more efficiently than the automatic enrolment scheme. It is an empirical question, therefore, which types of nudges actually do achieve the behavioural outcomes most effectively in each policy area; and, in cases where this is shown to be those that aim at securing moral compliance, the weighting issue will not arise. Furthermore, we have seen that one set of nonenforceable duties would not provide this type of choice between *Nudge A* and *Nudge B* either since, by their very nature, they cannot be fulfilled through certain measures, meaning that a conformity-driven System 1 nudge would not be available for the purpose of moral improvement in these cases.

There will certainly be cases, however, where a choice architect does have such a choice to make between policy designs and where there is data that a design which does not support morally responsible agency is more effective than its counterpart that does. Take the example of organ donation again. We might assume that a default rule that presumes consent for cadaveric organ donation is more effective at securing conformity with the duty to aid others (when this has little or no cost to oneself) than is an active choosing nudge design, though the latter are better at securing compliance.[147] In this type of case, it is possible that the *pro tanto* reasons in favour of active choosing will be outweighed by the considerations of those in society who are or will be in need of a life-saving organ transplant. Some governments have decided this way; others, including the UK, have not, and have adopted a prompted choice design that fits into my "inducing moral deliberation" category. It is not clear to me which all things considered judgement is correct in this case; this depends partly on consideration of deeper issues, such as

[147] Like many real-world cases, assessing whether this is in fact the case is made somewhat messy by the possibility that a proportion of those people who are now registered as organ donors under the automatic enrolment scheme would otherwise have opted in freely without any nudges or would have done so as a result of an active choosing nudge. Similarly, it is possible that the default might be responsible for encouraging some people to appreciate the reasons in favour of registering as an organ donor by, e.g., "conveying the impression that decision-makers should depart from the default only if they can identify good reasons to do so" (Hanna 2015: 628).

whether or not the duty to aid others in this way is subject to an injunction, meaning that the state cannot interfere coercively with a person's choices relating to organ donation (Kagan 1989: 219-230), which would rule out those System 1 nudges that count as *coercion-as-involuntariness*. I do not take a position on this issue, and remain open to the idea that an all things considered judgement in this and other similar cases would see the reasons in favour of supporting compliance outweighed by the requirement for harm-prevention.

But we can also think of examples in which the reverse might be the case, that is, where reasons for compliance outweigh reasons for conformity. As Clayton and Moles (forthcoming) contend, we need to recognise that "individuals can have a valid complaint against others reasoning in the wrong way as well as behaving in the wrong way and that the latter complaint does not always defeat the former". Consider a nudge that is designed to reduce implicit racial biases in hiring practices that has the side-effect of reducing in some way the target's ability to evaluate candidates. Her decisions are now not motivated by implicit racist attitudes, but she does fail more often in hiring the most qualified candidates.[148] Adopting this type of policy may lead to lesser conformity with the principle of fairness in hiring, but it may represent the morally desirable action all things considered, given the wrongness attendant on racial disadvantage and its (implicit) proliferation and, to a lesser degree, the ethical value to the target of not acting in a way that she would reject on the basis of moral beliefs she has.[149] This would give us reasons to select *Nudge B* over *Nudge A*.

Speaking generally, there are also likely to be cases in which citizens' interests (and their collective interest) in living in a political community that respects their claim to be treated as moral agents, capable of living up to the demands of morality for the right reasons irrespective of their potential moral deficiency, outweighs the potential welfare losses that might come with choosing *Nudge B* over *Nudge A*. Citizens might acknowledge that this *respect-based reason* means that they will not always receive the treatment they are owed from others, but nonetheless wish to live in a political community that respects theirs' and others' claim against it *not* to disable their ability to respond to reasons about what they have reason to do – by, e.g., using psychologically manipulative nudges to ensure higher rates of conformity with nonenforceable duties – unless the stakes are sufficiently high. This respect-based reason does not, however, seem to preclude the government from taking certain actions to create or sustain an

[148] This example is a very lightly modified version of one offered in Clayton and Moles (forthcoming).
[149] Here I assume that the target does not explicitly hold racist beliefs.

environment that offers some support to citizens with regards to responding to the reasons that apply to them. Given the threat to our ability to respond to reasons that comes from automaticity and its associated cognitive biases, compliance-supporting nudges can be viewed as respecting, rather than disabling, this ability – and therefore as respecting citizens' claim to be treated as a moral agent.[150]

The account of moral nudging I have defended seeks to respect this claim of citizens, and provides a sketch of some practicable policies that represent the view that:

> "[t]he more we are able to understand and correctly and sensitively evaluate our world, the more responsible we are able to be in acting within and upon it. Insofar as we want to promote freedom and responsibility, then, […] we can do so by promoting as well as by exercising faculties of reason, perception, and reflection, by encouraging as well as by cultivating open and active minds and attitudes of alertness and sensitivity to the world. These are what we need if we are to have the ability and the freedom to see things aright." (Wolf 2008: 274)

I do not profess to have assessed all of the relevant considerations relating to moral nudging. But this chapter has provided the first full account of the moralising potential of PCA and its permissibility, and has done so in a way that builds on the assessment of (purely) prudential nudging outlined in Chapter 4 in order to offer a complete theory of the political morality of public nudging.

§5.5: Conclusion

This chapter has examined the reasons that a political community has to nudge citizens for the sake of others – a complex analysis that has taken into account both the moral value and the potential ethical value of these kinds of policies. It has argued that, with regards to enforceable duties, there may be some role for (System 1) moral nudges, but that it is often be morally desirable to use traditional coercive strategies. It set out a conditional set of comparative conditions under which it would be permissible to use

[150] An analogous issue is raised in this moral case to that in the case of prudential nudging regarding whether or not the policies that have been defended are objectionably paternalistic. My account of moral nudging can be understood as a (weak) form of moral paternalism, since it defends certain kinds of public nudging on the grounds that (1) these interventions can help to bring about morally improved states, and that (2) certain such states are good for the individual in question (Dworkin 2005: 308). In my view, since (2) rests on the value of autonomy to living well, it is not objectionably paternalistic (i.e., disrespectful) for the same reasons as those outlined in §4.4.4 in defence of autonomy-supporting nudge paternalism. In the case of moral nudging, which is done for the sake of others, this ethical value is not what principally motivates these policies; but it becomes particularly relevant to the analysis because the main promise of PCA comes not from enforcing morality, but from helping governments to deal with problems relating to nonenforceability.

moral nudges as a means of securing moral conformity. The principal promise of moral choice architecture has been shown to come from the role that PCA can play with bringing about greater compliance with respect to nonenforceable duties. To this end, the chapter has outlined three permissible types of moral nudge that are justified by reference to their ability both to prevent harm to others and to support autonomy in moral action, which has ethical value for the agent. Such nudges count as moral nudges because of the type of behaviour they target; but they might also be understood as representing a distinctive type of prudential nudging. This fulfils my ambition to present a complete, yet simple account of permissible public nudging.

6. Permissible Public Nudging in Practice

The preceding chapters, comprising the main argumentative body of the thesis have set out a more complete and philosophically sophisticated theory of the political morality of nudging than is currently present in the literature. This final substantive chapter moves on to the practical question of what this extended argument can and does say about the way in which governments should make behavioural policies today. After discussing the relationship between philosophy and public policy (§6.1), the chapter draws out the essence of the general argument offered in this thesis and elucidates the two central ethical principles of the *transformative nudge model* (§6.2). It then sets out how this model might be applied to policy, particularly within the UK's Behavioural Insights Team, and discusses some implementation issues (§6.3). The chapter rounds up by reviewing some of the limitations of my argument in its current form and, in doing so, sets out an agenda for future research (§6.4).

§6.1: A Note on Political Philosophy and Public Policy

Two points are worth making at the outset of this chapter relating to the relationship between political philosophy and public policy. These are important because they have implications both for what we might reasonably expect in terms of policy prescriptions and for how we should understand the value of philosophical contributions to the policy debate.

The first is the acknowledgement that "philosophy, even political philosophy, has only a modest and specific part to play when it comes to policy prescriptions" (Brighouse and Swift 2014: 176). The normative considerations we have focused on in this thesis are but one kind of consideration at stake in policy debates, meaning that these principled considerations need to be examined alongside a set of complex empirical considerations. This amounts to recognising the (widely accepted) claim that good policies need to be sufficiently attentive to the relevant non-normative facts, in addition to being grounded on sound principles – and that the role of philosophers and political theorists is chiefly to contribute to the latter enterprise. As a result, this sets the ambitions of the chapter as outlining the essence of the extended argument developed in these pages in a way that is accessible to nonspecialists and presenting ethical guidelines for designing behavioural policies that follow from this philosophical account of permissible nudging. This is to say, more negatively, that it will not be in a position to

provide a complete evaluation of the considerations at stake in the policy debate over PCA.

Second, at the same time, it is crucial not to underplay the importance of this role for political philosophy, or to allow it to be significantly underplayed or misrepresented within this policy debate. One of the main motivations for this thesis is to address the fact that such an undervaluation and misrepresentation has occurred in this case, and to highlight and to seek to remedy the problematic effects that this has had on how the debate has developed. More specifically, as has been mentioned before, nudging's main proponent is suspicious about the role of theory. Not wanting to fall into what he calls the "trap of abstraction", Sunstein has bracketed many of the serious philosophical questions that arise, and has sought to gain justificatory purchase for PCA in ethically non-committal, or at least ecumenical, ways. His approach is marked by the thought is that people can agree that nudges are to be welcomed, or at least that a significant proportion of them are, despite having disagreements about foundational questions in moral and political philosophy (Sunstein 2016a: 52).

There are two points to make in response to this approach. First, as plausible as it sounds, this methodological approach can work only if disagreements about such foundational questions do not in fact matter for the particular controversies under examination. Sunstein's "hope" is that "those disagreements will not arise often" (ibid.). This thesis has shown that this viewpoint is overly optimistic. Second, Sunstein's approach appears to consider it sufficient that, generally, people agree that certain nudges are welcome. Yet such agreement is insufficient, since what we need to show, instead, is that it is morally permissible to impose a nudge on a particular person who objects. In other words, it is the relationship between the political community and the individual that needs justifying, and this cannot be done merely by establishing that generally people agree to this particular kind of relationship. What needs establishing is that the individual in question has no valid complaint against the nudge in question being imposed upon her. This is why it is necessary to engage with, rather than bracket, a number of the philosophical questions raised by the use of PCA, and to offer a theory of the political morality of nudging.

Sunstein recently offered his first sustained attempt to attend to the normative issues raised by choice architecture (Sunstein 2016a). This was published well over a decade after the original statement of the pro-nudge argument (in Sunstein and Thaler 2003) and was only deemed necessary, it might be assumed, due to ongoing critical

engagement with their earlier, less developed arguments. Nonetheless, as one reviewer of *The Ethics of Influence* writes:

> "The result is an excellent primer of the relevant debates on nudges... But we do not get a bona fide philosophical inquiry into the underlying ethical issues – one that would [plough] headlong into the most difficult dilemmas, rather than immediately veering onto safe ground. And this is the sort of investigation that is sorely needed." (Johnson 2016)

Another reviewer writes that his "pragmatist avoidance of issues of principle frequently conspire to muddy the conceptual waters in troubling ways" (Anderson 2010: 375). These two criticisms represent precisely where the "modest and specific" function of political philosophers comes into play. This thesis has provided this "sorely needed" conceptual and normative analysis and, in doing so, has uncovered several of the weaknesses of a pragmatist methodology. Many of these comes as a result of failing adequately to delineate the relevant conceptual and philosophical considerations, which has led both to a lack of clarity in the foregoing discussion and to ethical analysis that attempts to deal with these confusions in an unsatisfactorily ad hoc fashion.

Even if the policy prescriptions of the two views (i.e., Sunstein's view and the theory developed here) were identical, an assumption we can grant for the sake of argument at this point, the former would lack the justificatory force available to the latter. This is because the arguments offered for these prescriptions by the former can reasonably be expected to be either incomplete or mistaken, or both. In the first instance, the defences offered by Sunstein might be *incomplete* for three possible reasons, that is, because the normative considerations they appeal to (i) are not sufficiently deep, (ii) fail to address important objections, or (iii) are insufficiently nuanced. In the second instance, such defences might be *mistaken* because the arguments offered for them are not sound, where soundness requires (i) validity (i.e., correct inferences) and (ii) true premises. As I have argued in the preceding chapters, the standard pro-nudge argument falls foul in both of these ways. It is incomplete because, among other things, it fails to acknowledge the morally relevant differences between nudging in the prudential and moral domains. And it is mistaken because, among other things, its account of well-being lacks plausibility. As a result, Sunstein's arguments provide insufficient normative justification for the (stipulated) policy prescriptions, assuming that my arguments are accepted.

The implication of this is that the argument presented in these pages, or something like it that also commits to performing the specific functions assigned to political philosophy, is necessary for grounding the permissible use of nudges, even in the unlikely event that it had no distinct prescriptions for policymaking in practice (for a similar point, see Dworkin 2011: 110). Hence, even while accepting the limits of the role of political theory, we should be clear about the value of its specific contributions, both clarificatory and normative, to the debate over the ethics of nudge policy.[151]

§6.2: From Extended Argument to Ethical Principles

The two parts of the extended argument – prudential and moral – were carried out separately in order to keep important differences from clouding our ethical analyses. Nevertheless, amid the specificity and complexities of these detailed normative examinations, the two analyses share an argumentative essence and this provides a basis on which to develop overarching ethical guidelines for the design of nudge policies. This section begins to move us from political theory to political practice by, first, extracting and expressing the essential form of the thesis's extended argument and, second, encapsulating the principles that, according to this argument, should be adopted by policymakers when designing PCA.

§6.2.1: Essence of Principled Evaluation: Agency-Enhancement

The account of living well developed in Chapter 3 has played a central role in directing our normative investigation. It claims that, for an individual to live her life well, she must live it in an active sense: prudentially, her life goes better the more she is the (part) author of it; morally, her life goes better the more she fulfils her moral duties to others out of a recognition of moral reasons. Ultimately, both rest on the value of personal autonomy to a person's life. Hence, in so far as nudges can be justified on first-personal "welfare-promotion" grounds, to use Thaler and Sunstein's terms, such governmental influences should be supportive of the development and exercise of these capacities for autonomous and morally responsible agency.

Across both strands of normative analysis, then, the kinds of psychologically-informed behavioural interventions that have been deemed permissible share a similar character: they are compatible with creating and sustaining the conditions for living well on account of their capacity to play what we might call an "ecological-educative" role in

[151] See Wolff (2011) for more on the role of political philosophers in practical applied policy debates.

supporting citizens' agency.[152] Those interventions that are designed to act on people in other ways, such as to modify their behaviour "by stealth rather than engagement" (Rowson 2011: 16), fail generally to meet the conditions for permissibility.[153] Once we acknowledge that the government is inevitably involved in structuring the landscape of choice, policymakers must use this power (when they do) not as a means for modifying or controlling their citizens' behaviour; governments have other ways of doing this that are safeguarded by a transparency that comes via the democratic processes attendant on legal and regulatory action. On the occasions when they do intentionally make use of this power (recall that "choice architecture" is not inevitable, even though the government is inevitably involved in structuring "the landscape of choice"), they must do so carefully, and in ways that aim at supporting citizens' agency to act in autonomous and morally responsible ways. This is because there is an important difference between a government employing nudge techniques to bring about what we might call "first order change" (i.e., behaviour modification), and its doing so for the sake of supporting the possibility of "second order change" (i.e., agency-enhancement, which may lead to the same targeted behaviour modification being decided upon and performed by the agent herself) (Watzlawick, Weakland, and Fisch 1974).

This argument directly challenges the claim, proffered by a number of people in the literature and succinctly captured by Rowson's formulation, that the "deepest problem with nudge is that it is not transformative" (2011: 16; see also Goodwin 2012: 86). By this, they mean something like that nudges are not able to help foster the types of human learning and development processes that can contribute to more enduring changes to our behaviour or our beliefs. As the reader will now know, my argument is in agreement with this critique to a certain extent: in both the prudential and moral domains, some kinds of nudges are ruled out as impermissible for precisely this reason. But, importantly, the thesis has distinguished those types of PCA that are unable to promote this type of development and agency-enhancement – we can call these "non-transformative nudges" to adopt the terminology used by the critique – from a set that are able to play a role in these processes – "transformative nudges". Hence, the model developed here can be labelled the *transformative nudge model*.

---

[152] For more on what is meant by nudging's "ecological-educative" capacity, see §6.2.3.

[153] The inclusion of the word 'generally' here is to acknowledge that the case is slightly more complicated in the moral domain, as we saw in Chapter 5. In §5.3, I outlined some conditions under which it would be all things considered morally permissible to use psychologically manipulative moral nudges, whilst simultaneously arguing that these conditions are unlikely to be met on very many occasions, since (additional) coercive measures may sometimes or often be morally preferable in these kinds of cases.

§6.2.2: The Transformative Nudge Model: Two Ethical Principles

The essence of the transformative nudge model, then, is agency-enhancement. Recall the distinction drawn between "System 1" and "System 2" nudges: in brief, the former bring about behavioural modifications via nonconscious automatic processes, while the latter act on us in ways that encourage conscious, deliberative choices. Across both the prudential and moral analyses, it has been shown that at least one form of each type of nudging is permitted. Following on from this, two main normative principles form the basis of the ethical framework for designing permissible nudge policies: (1) the *nudge-to-deliberate principle*, and (2) the *nudge-to-motivate principle*. The permissible System 2 nudges relate to the former, while the permissible System 1 nudges relate to the latter. I will briefly explicate each principle in turn.

The first principle of the transformative nudge model is the *nudge-to-deliberate principle*. It maintains that PCA is morally permissible in so far as it supports people's practical reasoning and autonomous decision-making. According to this principle nudges can be used to activate or engage citizens' deliberative thinking processes in contexts in which their behaviour is likely to be subject to a predictable cognitive bias or to predictable motivational or attentional strain.[154] In doing so, it will often act as a form of *psychological debiasing* and, in some cases, also as a *counter-nudge*. Much of the world in which we navigate our lives is not set up "neutrally", as it were. We face a near-constant stream of information and influences seeking to shape our preferences, direct and demand our attention, and steer our behavioural responses. Governments can protect and support their citizens' exercise of autonomous agency, and support them in freely complying with their moral duties, by using nudges that operate via the *nudge-autonomy* and *inducing moral deliberation* mechanisms, respectively, since these mechanisms involve enhancing in some way an individual's general ability to engage in deliberation relevant to the situation she is in (the *deliberation-activating view*). The *public-ecological persuasion* mechanism discussed in the context of moral nudging also falls under the nudge-to-deliberate principle (see the top row of Table 1, below), but it differs from these other

[154] I do not want the terminology of "deliberation" in the nudge-to-deliberate principle to misrepresent the principle's aim. The nudges that fall into this category need not engage their target's deliberative processes in ways that are cognitively burdensome. In many cases, it will involve only the momentary conscious process of 'checking' – which may result in the individual confirming the automatic response or replacing it with a behaviour that is more in line with her reflected preferences, neither of which require any significant cognitive load. In such cases, the term "deliberate" possibly gives a false sense of the type of cognitive processes that are triggered, given its association with long, careful consideration. There will also be some cases in which this more cognitively burdensome consideration is triggered by the PCA, especially in the case of moral nudging. Whether or not this is the case will often depend on whether a particular individual already has a reflected preference on the matter at hand or not. If not, it may encourage her to consider her view, whereas it may only involve a momentary checking otherwise.

two mechanisms in so far as it seeks not only to activate deliberation, but to do so in a directive way, i.e., to steer people towards recognition of a particular reason. Hence, this mechanism has the additional aim of enhancing an individual's ability to track the truth about what reasons apply to her (the *truth-tracking view*).[155]

	Prudential	*Moral **
"System 2"	Nudge-autonomy	Inducing moral deliberation; Public-ecological persuasion
"System 1"	Auto-nudge	Motivational scaffolding

Table 1: A typology of permissible nudge mechanisms captured by the two principles (the top row represents the nudge-to-deliberate principle across both domains; the bottom row the nudge-to-motivate principle).

The second principle, the *nudge-to-motivate principle*, holds that nudges are permissible in so far as they help people to act in accordance with their own beliefs and intentions (see the bottom row of Table 1). In cases where an agent voluntarily opts in to receiving motivational support from PCA, such *auto-nudges* count as ways of supporting and enhancing her autonomy by ensuring that her future behaviour is constrained so as to align with her goals and values. In the moral realm, *motivational scaffolding* encompasses more than (explicitly) consented-to System 1 nudges. It also includes nudges that reduce implicit biases, which highlights that interventions that make use of and seek to (re)shape our automatic responses are not necessarily problematic for autonomous and morally responsible agency.

The table above is structured along the lines drawn by the thesis, and thereby categorises the nudge mechanisms relating to whether the intervention is principally for the sake of the individual upon which the nudge is acting or for the sake of others. The moral column is labelled *Moral ** because, on my view, although the *content* of the agency-enhancing nudges in this column is moral (i.e., for the sake of others), prudential reasons play an important role in motivating and justifying these mechanisms. That is, the moral nudges that are justified by the nudge-to-deliberate and nudge-to-motivate principles are partially justified by the idea that it is in the interests of the target of these interventions to act on one's own judgement; in particular, one's life goes better when

[155] It should also be added that an implicit (empirical) assumption of the other mechanisms that fall within the nudge-to-deliberate principle is that deliberation-activation will tend also to enhance a person's ability to truth-track with regards to the reasons that apply to her; but this need not be the case.

one is sensitive to the ways in which one's behaviour and attitudes harm or wrong others, and when the appreciation of this consideration motivates one to act in accordance with moral reasons.

Together, these two principles should direct the design of PCA within those societies that currently employ this policy lever and those that may want to do so in the future. When used in accordance with these principles, nudges can be used by governments as a means of enhancing citizens' agency in ways that support the recovery of control over their thinking and behaviour. Any nudges that do not comply with one of these two principles – with the exception of the category of moral nudges that might be more effective than coercive measures at bringing about conformity with enforceable duties – are ruled out as impermissible.

§6.2.3: Nudging's "Ecological-Educative" Capacity

Others who have advanced broadly similar kinds of arguments to mine, at least in so far as they might also be categorised as "liberal" responses to the standard nudge view, have set up their alternative proposals in (terminological) opposition to nudging. Such views are grounded in the view that nudging is problematic because "it is not transformative" (Rowson 2011: 16) or, in other words, because it lacks the capacity to play any educative role. There are two main examples of such alternative "liberal-educative" proposals in the literature: the 'think' model advanced by Peter John et al. (2009, 2013) and the 'risk literacy' model grounded in the work of psychologist Gerd Gigerenzer (2014, 2015). It is instructive to offer a brief overview of these models in order to highlight both the similarities with and differences from the transformative nudge model. Doing so will, I hope, make clearer what is distinctive about this model vis-à-vis other liberal-educative alternatives and, consequently, what is meant by the "ecological-educative" capacity that I ascribe to some forms of choice architecture. It will also go some way towards answering the terminological question about my decision to stick with the nudge vocabulary, when others have abandoned it.

'Think' refers to a behaviour change strategy that takes its lead from deliberative theories of democracy. The group of political scientists behind it begin with the thought that citizens are "reasonable, knowledge hungry and capable of collective reflection" (John et al. 2009: 366). This assumption about human nature grounds the claim that, despite people having a series of cognitive biases that can cause them to choose in non-rational ways, social fora and institutional settings through which deliberation and free discussion take place provide a means by which citizens can overcome non-rational and

immoral or anti-social beliefs and behaviours. The thought, which follows from their underlying commitment to deliberative democracy, is that 'think' has both an epistemic and a moral dimension. First, deliberation is assumed to have an educational effect: providing citizens with a public space in which they can step away from day-to-day life and reflect on a wide range of social policy choices with others in their community enables them to increase their knowledge and understanding, often by learning from others. Second, it is thought also to have a "moralising effect" (Miller 1992): dialogue and discussion, where citizens are required to justify their perspectives and preferences to others, helps to reduce or eliminate non-rational and unreasonably self-interested preferences (John et al. 2009: 364).[156]

In contrast, 'risk literacy' is grounded in an interpretation of the empirical psychological evidence, rather than in democratic theory; although, as we will see, there are significant underlying similarities between the two models. On Gigerenzer's view, it is not "fast and frugal" heuristic-based reasoning that is the cause of the non-rational choices that behaviour change strategies wish to target (Gigerenzer and Goldstein 1996).[157] Rather, he diagnoses the cause as being that citizens are not currently receiving the types of education that enable them to develop the skills necessary for dealing with the risk and uncertainty pervasive in modern technological societies. These skills relate to the development of both automatic and deliberative cognitive processes. On this theory, practical rationality is comprised both of the "unconscious intelligence" provided by intuitive and automatic thinking, which is required for dealing with uncertainty (i.e., unknown risks), and the logical and statistical reasoning associated with deliberative thinking, which is required for dealing with known risks (Gigerenzer 2014: 123-124). What is needed, then, is to address the "amazing blind spot concerning risk literacy" in our current educational system (ibid.: 13-14).[158]

Developing risk literacy, and thus "risk savvy citizens", requires teaching a basic knowledge of humans' intuitive psychology and an understanding of statistical information – both of which can and should, according to this model, be taught to citizens in schools from a young age, and extend into higher education too (especially

[156] In his argument against nudging, Goodwin also suggests that "the government ought to embrace more deliberative models and public engagement, which arguably encourage people to think more collectively and engage with issues more deeply" (2012: 90). This shares elements with the 'think' model, in particular the aspects relating to collective deliberation.

[157] For more on the differences between Gigerenzer's "simple heuristic" program and Tversky and Kahneman's "heuristics and biases" program, see Grüne-Yanoff and Hertwig (2016).

[158] Gigerenzer (2010: 469) argues that: "We need to change school curricula. Our children learn the mathematics of certainty, such as geometry and trigonometry, but not the mathematics of uncertainty, that is, statistical thinking. Statistical literacy should be taught as early as reading and writing are."

for those attending medical and law schools). These educational reforms would focus on "training the individual to avoid the bias and, thereby, equipping him to maximize his own welfare without the help of some policy maker or other outside party" (Hertwig and Ryall, unpublished: 7). Such options, recently labelled "boosts" by Grüne-Yanoff and Hertwig (2016), aim to provide people with "the ability to independently surmount a bias for an entire class of problems – but, at the cost of an up-front resource investment" (Hertwig and Ryall, unpublished: 8).

These two models share with the transformative nudge model an ambition to bring about more lasting behavioural changes by supporting citizens' decision-making competencies and reasoning processes. In other words, they all have an educative core.[159] It is not clear beyond this whether the three broadly liberal-educative views share the same ends, since the 'think' and 'risk literacy' models specify their ends in fairly vague terms. For instance, it is not clear whether the goal of 'think' is personal autonomy, democratic competence or "good citizenship", civic renewal more generally, etc., or some mix of these. A different way of putting this is that the transformative nudge model follows from and is set within a full theory of political morality, which is not (currently) the case for the two alternatives. It is not necessarily the case, then, that 'think' and 'risk literacy' share the same ends but adopt different means to realising them; we would need to know more about their more specific aims to conclude this. And, despite being based (at least initially) on a rejection of the nudge programme, both of these models might be best understood as offering complementary proposals. As advocates of 'think' acknowledge in the mature statement of their view, this strategy is not offered in opposition to nudging, but rather provides an additional, complementary strategy that can "deepen and broaden out the behaviour-change programme" (John et al. 2013: 3). Similarly, the educational reforms advocated by 'risk literacy' – which include adding subjects like health literacy, financial literacy, and digital risk competence to school curricula so that students learn skills required for living in a "risk society" (Beck 1992) – can be pursued alongside nudges, as a complementary response to the psychological evidence aimed at those in (often mandatory) full-time education rather than (adult) citizens influenced by PCA.

These two models differ from the transformative nudge model, then, since the proposals offered by 'think' and 'risk literacy' are based in institutional and educational reform, rather than in the design of psychologically-informed behavioural public

[159] This is unsurprising. As Ben-Porath (2010: 9) states: "The two domains making up the landscape of choice – the political and the personal – converge in the realm of education."

policies. The distinctiveness of the transformative nudge model comes from its ecological-educative capacity: it seeks to enhance citizens' agency (like 'think' and 'risk literacy'), but to do so through the design of policy ecologies rather than institutional and educational reforms (like the standard, non-selective account of nudging). Of course, this second means-related aspect is unsurprising. This thesis examines the ethical dimensions of one particular means of intervention in order to answer the question of what forms of PCA are permissible in light of our new understanding of human action. The transformative nudge model is the answer that follows from the extended arguments presented here. A central point has been to show that liberals do not necessarily need to reject PCA as a means of influence out of hand and come up with alternative ways of responding to the psychological evidence. It is possible for nudges to play some role at least in achieving the liberal-educative goals that those who often dismiss nudging as non-transformative wish to achieve. We can hold this view while simultaneously acknowledging that alternatives such as 'think' and 'risk literacy' may: (i) complement transformative nudges, (ii) be deemed desirable or even necessary additional responses, and/or (iii) have advantages over the transformative nudge model.

This does not fully answer the semantic question that spurred this clarificatory detour. Part of the reason for retaining the language of nudging does follow from the point made above, namely, that by providing a comprehensive ethical assessment of the use of PCA the thesis has been able to show that some kinds of nudges are permissible (*contra* some liberal views). But there is another, perhaps implicit, dimension to the terminological question: why retain the overall language of nudging given the potential for confusion between its permissible and impermissible forms? I could, for instance, have labelled my positive view in non-nudge terms in order to avoid any such potential confusion. The main reason for resisting this is that adopting different terminology would concede too much to the standard account of nudging. Nudging is something that has won considerable favour with policymakers, and it has become or is becoming a part of the institutional mind-set or status quo in governments across the world. These institutional facts are important. To offer a different (non-"nudge") account would not only be unrepresentative of the project, it also would undermine the success of its main practical goal, which is to ensure, given that nudges are in fact a (fairly fixed) feature of contemporary governmentality, that there are ethical conditions placed on which kinds of nudges are used, through the establishment of ethical guidelines that at least approximate those outlined above. Hence, my analysis is well positioned both to

take advantage of the approval that nudging has generally won in policymaking circles and to address the underlying reservations, also present within these circles (e.g., House of Lords 2011) as well as in academia and the media, about whether or not nudges raise ethical qualms.

This point extends to a pragmatic advantage that the transformative nudge model has over the two alternatives outlined above. As noted earlier in the thesis, much of this support from policymakers comes from nudging's *low cost*, both financially and politically (since it is, in general, a non-regulatory policy lever). The transformative nudging model does not need to gather any additional political will or support in the way that 'think' and 'risk literacy' would need to. It suggests a modification to existing practices (and one that, as we will see in the next section, aligns with the public's views on the acceptability of these techniques and which would, therefore, have political benefit), rather than the introduction of any costly new policies and reforms. Strategically, then, these favourable practical features make it worth pursuing as one response (amid others) to the psychological evidence. The next section addresses both the implications of this proposal for nudge policy in practice.

§6.3: Transformative Nudging in Practice

The next task for this chapter is to try to translate the transformative nudge model and its two ethical guidelines into an accessible policy framework that can be used by governments – or at least the early stages of one.

§6.3.1: The Beginnings of an Ethically-Sensitive Policy Framework

What difference might the ethical principles make to existing governmental approaches to designing PCA here and now? One way of answering this question would be to engage in comparative analysis of different ethical guidelines – the transformative nudge model versus the ethical guidelines currently in use by, e.g., the Behavioural Insights Team (BIT).[160] However, to the best of my knowledge, the behavioural insights teams working across the UK government do not have any specific, formalised ethical

[160] I use the example of the BIT throughout this subsection for a number of reasons. Most pragmatically from my perspective, the BIT is the behavioural policy unit in my own political community, the United Kingdom. More generally, though, it was the first such unit to be set up and is "the model that many other governments have begun to follow" (Halpern and Sanders 2016: 54). If the BIT made changes to its practices, this would certainly have wide ranging implications on behavioural policymaking practices generally.

guidelines that are used for intervention design.[161] It seems that the ethical concern is directed solely towards the issue of BIT's trialling, usually without citizens' consent, groups of the population for the randomised control trials that are the focus of the team's work. A more formal "trial protocol clearance process" has recently been put in place for this purpose (Halpern 2015a: 311). Even in the House of Lords investigation into the government's use of behaviour change techniques, consideration of the ethical dimensions of the issue was relegated to an appendix (House of Lords 2011: 105-109). Unsurprisingly, then, building such guidelines into the framework(s) currently employed by the behavioural insights teams is the main implication for behavioural policymaking argued for here. This subsection seeks to engage in a deeper fashion with current practices (and the ways in which these are framed and discussed by policy designers) in order to explore how formally establishing the ethical principles outlined above would affect the operation of the BIT.

Let's begin by briefly outlining the policy frameworks provided by the BIT in recent years in order to get a better sense of current practices and to highlight the current level of consideration of the ethical dimensions. The BIT's first framework was outlined in the report 'MINDSPACE: Influencing Behaviour through Public Policy' (Dolan et al. 2010). Its title is a mnemonic designed to be used as "a checklist for policy-makers", listing the main influences on people's behaviour which might be taken into account when making policies (ibid.: 7-8).[162] In order to build these into policy-making, the authors proposed the "6 Es Framework": Explore, Enable, Encourage, Engage, Exemplify, and Evaluate (ibid.: 49-62). As Halpern recounts in his book *Inside the Nudge Unit*, after about a year or so a simplified framework was developed (2015a: 60). It was replaced by the more straightforward mnemonic EAST, to capture the BIT's four key principles for applying behavioural insights to policy interventions: Easy, Attractive, Social, and Timely (Service et al. 2014; Halpern 2015a: 59-149). The first

[161] There are some ethical constraints on what can be done, but these are set out by the particular department or agency with which the behavioural insight team is working, not by the BIT itself. For example, those working in Public Health England's BIT do have fairly rigorous ethical procedures in place, but "only insofar as what's necessary for passing NHS ethics. Obviously there's no point in designing a trial to only have it rejected by the ethics board, so the trial will be designed with reference to what the NHS ethics committee permits" (personal correspondence with an employee at the BIT).

[162] These are: Messenger (we are heavily influenced by who communicates information), Incentives (our responses to incentives are shaped by predictable mental shortcuts), Norms (we are strongly influenced by what others do), Defaults (we 'go with the flow' of pre-set options), Salience (our attention is drawn to what is novel and seems relevant to us), Priming (our acts are often influenced by sub-conscious cues), Affect (our emotional associations can powerfully shape our actions), Commitments (we seek to be consistent with our public promises and reciprocate acts), and Ego (we act in ways that make us feel better about ourselves) (Dolan et al. 2010: 8; see pp. 18-48 for more on this framework).

principle – Make it Easy – suggests harnessing the power of defaults, reducing the "hassle factor" of taking up a service, and simplifying messages. The second – Make it Attractive – comprises of attracting attention and designing rewards and sanctions for maximum effect. The third – Make it Social – captures the results of showing that most people perform the desired behaviour, using the power of networks, and encouraging people to make a commitment to others. Lastly, the fourth – Make it Timely – recommends prompting people when they are likely to be most receptive and helping people plan their response to events (Service et al. 2014: 9-42). Some of these evidence-based recommendations might be deemed permissible according to the transformative nudge model's ethical principles; clearly, others would not. For now, I leave aside the exploration of these finer details and instead address the more general question of why this framework is silent on the normative dimensions of behavioural policymaking, and how it might be possible to integrate the relevant normative considerations into this existing framework.

There are a number of factors explaining why the BIT has developed and found favour in the way it has, many of which are contingent political-strategical ones (see John 2014). One of the main reasons, according to Peter John, is that it has been able "to promote a more entrepreneurial approach to government by using randomized controlled trials (RCTs) as a robust method of policy evaluation" (2014: 257). It thereby promises to deliver more effective and cheaper behaviour change interventions by using evidence from the cognitive and behavioural sciences to design trial interventions and to test these in RCTs to generate data about "what works" (Halpern 2015a: 281-294; for a more critical discussion of RCTs, see Cartwright 2010). This extends the trending category of "evidence-based policy". Indeed, according to the BIT's Chief Executive, it takes this approach to the next level, because it represents "the rise of experimental government" (Halpern 2015a: 295). The EAST framework is embedded within this broader methodology, and is comprised of four main stages (Figure 2, below).

These stages would all still be a part of any updated policy framework based on the normative arguments of this thesis; they represent the cumulative wisdom of behavioural policy experts over the years since (and indeed before) the BIT was established. But, it is nevertheless importantly *incomplete*. It does not recognise that behavioural policy design and decision making involves value judgements. As a result, it lacks any attempt to address the issue of how we should think about the relationship between normative and non-normative considerations in the design of public policy.

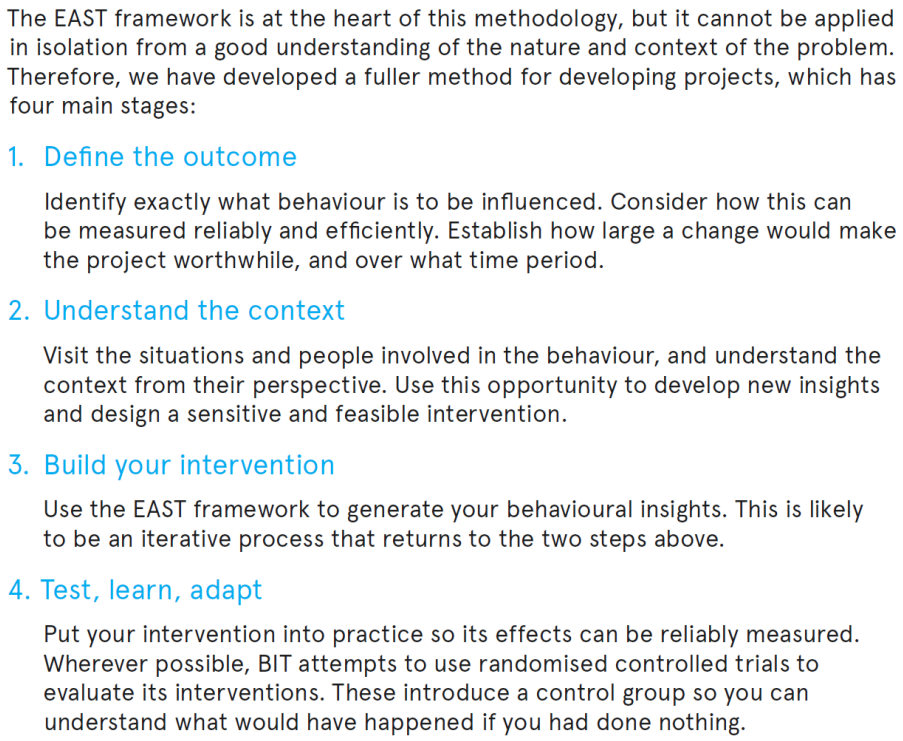

Figure 2: EAST's overarching experimental methodology (Service et al. 2014: 7).

How might we proceed in addressing this incompleteness? I am in agreement with Goodin's statement that, "Empirical and ethical theory ought both to be used, and used in tandem, to guide public policymaking" (Goodin 1982: 4). Yet it remains a complex task to specify precisely how this "in tandem" component would work in practice. One recent suggestion for thinking about how this might occur in educational policy has been offered by Brighouse et al. (2016). This also sets out a four-part procedural framework (see Figure 3, below).

The main aim of this framework is to aid policymakers in making decisions that "explicitly combine values and evidence" (Brighouse et al. 2016: 20). The particular framework mapped out below is directed at a specific policy area; nonetheless, the procedure it outlines has wider relevance. It provides us with a general approach to thinking about how this type of framework might be integrated with BIT's existing methodology outlined in Figure 2. It is worth noting here that the authors clarify that, "Although we describe the procedure in terms of ordered steps, it is better to think of the parts simply as distinct elements of the process, all of which need to be addressed, often iteratively and not necessarily in the sequence presented here" (ibid.).

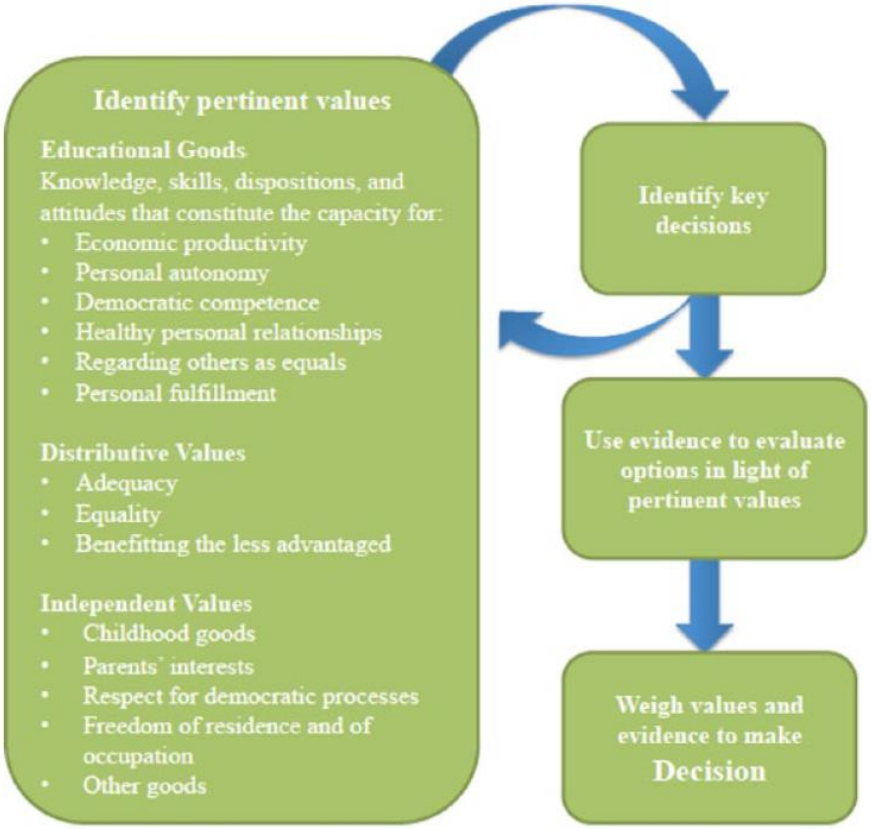

Figure 3: Brighouse et al.'s framework for decision-making about educational policy, which suggests how to incorporate *educational* goods and values into standard policy decision-making procedures (2016: 20).

In designing permissible PCA, we might start with identifying a particular policy problem or goal that might have a role for nudges to play in bringing about positive behavioural outcomes. This initial stage is best captured by the BIT method's first stage, 'Defining the outcome'. We might then move on to identifying the pertinent values in play – the first stage of Brighouse et al.'s model and the main contribution of this thesis. On my account, the main values are living well, personal autonomy, harm-prevention, and morally responsible agency. But, there may be other values in play either generally or within the particular domain to which the policy goal in question belongs, and consequently there may be tensions between values (for more on this in the case of transformative nudges, see §6.3.2). We might then move on to identifying the potential choice-architectural means that could be used to more effectively achieve this policy goal, while not offending against (and hopefully furthering) the values outlined in the preceding stage. This requires "contextual work" (Brighouse et al. 2016: 20) and so takes the form of being a hybrid of the BIT method's second stage, 'Understand the

context', and Brighouse et al.'s second part of their procedure. Sometimes there will be a number of different feasible options; other times there may only be one nudge mechanism that is able to meet these practical and ethical conditions. Where there are different options, we might design and test trial interventions to gather evidence about the relative effectiveness of these permissible options. This relates to Brighouse et al.'s third part of the procedure, about which the authors write: "The values guide the search for evidence, and the evidence makes possible the evaluation of the options" (2016: 21). Similarly, with respect to the design of nudge interventions, the ethical guidelines place constraints on the types of interventions that are designed and trialled in RCTs in the first place, and therefore on the kinds of evidence that is taken into account. (Note that, in terms of the BIT method's third stage, the full EAST framework would not be used as some of its effectiveness-based recommendations would not pass the standards generated by the two agency-enhancing ethical principles, the nudge-to-deliberate and nudge-to-motivate principles.) The final stage would be to decide whether or not to implement the public nudge on the available evidence of its effectiveness at achieving the policy goal and the attendant value of agency-enhancement, as well as how this weighs against other values identified in the earlier stages.

To the best of my knowledge, this is the first attempt to integrate existing experimental behavioural policy frameworks with philosophically well-defined ethical guidelines. My main hope is that the clearer understanding of the ethical and moral dimensions of nudging presented in these pages will be an important contribution to the two-way discussion that needs to happen between ethicists and policy practitioners about what using empirical and normative theory "in tandem" to guide the design of permissible public nudges might look like.[163] This is an important aim and (potential) real-world contribution because, up until now, my reading of the debate is that these two groups have been talking past one another, essentially having separate discussions.

Finally, I would like briefly to highlight something that I take to offer a source of optimism for this shared, collaborative project. Although the BIT was established by the Coalition government in 2010, this behavioural-based policy entrepreneurship in the UK can be traced back at least as far as a 2004 report published by the Cabinet Office's

[163] Policy practitioners are themselves becoming increasingly aware of the need to open up this discussion. In personal correspondence, one of the BIT's employees stated that: "The ethical issues surrounding these techniques are becoming increasingly prevalent in discussions and everyone in my team [the Behavioural Insights Team at Public Health England] is very interested in this. If you developed comprehensive, useable criteria that delineate what techniques are considered ethical/unethical and under what circumstances, there would be a lot of interest in a document of this kind!" Similar sentiments have been shared with me from other practitioners, including at the World Bank.

Strategy Unit under Labour – and written by David Halpern (the now Chief Executive of the BIT) and Geoff Mulgan (the now Chief Executive of Nesta, the organisation in which the BIT is housed), among others – called 'Personal Responsibility and Changing Behaviour: The State of Knowledge and its Implications for Public Policy'. In addition to making points about the cost-effectiveness of these types of interventions relative to "traditional service delivery", this report argues that there are good reasons to think about how behavioural strategies might be used for the sake of developing citizens' personal responsibility. The authors claim that:

> "There are strong moral and political arguments for protecting and enhancing personal responsibility. Most of the dominant traditions of social and political thought in the UK value individuals' and communities' ability to take control and act in their own best interests as goods in themselves. Other things being equal, they see it as better for governments to empower citizens as much as possible rather than making decisions on their behalf." (Halpern et al. 2004: 3)

Similarly, in one of the first official BIT publications, the MINDSPACE report six years later, personal responsibility remains the main focus of the relatively limited normative discussion, in which the authors mention that it is possible "for government to just supply the trigger or support for individuals to take greater personal responsibility" (Dolan et al. 2010: 70-71). These references, especially the earlier extended quotation from before the publication of *Nudge*, highlight that the original or underlying ambitions of the BIT policy entrepreneurs are perhaps more aligned with the transformative nudge model than we might have thought, especially if what they mean by "personal responsibility" can be understood primarily in terms of what we might think of as the ethical responsibility (or autonomy) discussed in Chapter 4, and moral responsibility discussed in Chapter 5.

The publication of *Nudge* a number of years later brought behavioural policy-making to the fore and laid the foundation for the creation of the BIT. But it also brought with it a way of framing and discussing the issues that has created a kind of conceptual and argumentative path dependency. The transformative nudge model recommends a return by BIT policy entrepreneurs to their earlier orientation and, with it, a recovery of their initial motivation for engaging with the evidence from the cognitive and behavioural sciences. The promise of this possibility gives me some hope, even if overly optimistic perhaps, that the model's ethical principles might be able to have some influence on the BIT's policy agenda as it develops into the future.

§6.3.2: Some Issues Relating to Implementation

Nevertheless, consideration of the implementation of the transformative nudge model does raise a host of complex issues – some practical in nature, others relating to tensions between values. As Brighouse et al. maintain with respect to the latter: "good policy making requires awareness of how decisions are likely to affect the full range of values at stake", which itself necessitates explicit and careful consideration of these tensions (2016: 4). Although such a complex, all things considered assessment is not something that can be delivered in this chapter (as I noted at the outset of this chapter), it is important carefully to delineate, formulate, and discuss these considerations. To this end, this section outlines and briefly addresses four issues that may be relevant to the design of nudge policies that have not yet been discussed. With no pretensions to exhaustiveness, the issues covered here relate to: (i) political feasibility; (ii) public acceptability; (iii) efficacy and superficiality; and (iv) distributive concerns.

First, as is the case with all policy levers, PCA is designed and delivered by human choice architects and their use is authorized by political decision-makers whose benevolence and competence cannot be assumed. It is necessary, therefore, to assess the political feasibility of institutionalising the transformative nudge model, including vis-à-vis other means of achieving its desired ends. Of course, this implementation issue also arises for the standard nudge view – and more so, it seems fair to say. One of the main practical reasons for carrying out this investigation is to set and establish ethical standards by which benevolence, or the lack of it, can be measured in some way and to design easy-to-use guidelines that reduce the risk that incompetence, in ethical reasoning or otherwise, might cause problems for the design of permissible behavioural policies.

Second, there is an issue about in what way(s), if at all, citizens' views about nudges matter in deciding whether PCA is morally permissible. Increasingly, empirical research is being carried out into public attitudes towards nudging, including how people distinguish among different types of nudges and what effect these perceived differences have on their views about the acceptability of these interventions. Some have already discussed the issue of what role such data should (or should not) play, and for what reasons (e.g., Sunstein 2016a: 116-120; Dolan et al. 2010: 63-65). I broadly agree with these existing views, which state that taking account of public attitudes is important in liberal-democratic societies, but that these data cannot play a decisive role in evaluating moral permissibility. The first part of this statement can be grounded by reference to different values (legitimacy, self-government, political autonomy, well-

being, etc.), but the general point is that democratic governments ought to attend to what citizens think of nudges.[164] However one chooses to make this first point, it seems that they can all agree that "it is not necessary to make strong claims about the wisdom of crowds […] in order to believe that an ethical judgment, on the part of those who might be subject to nudges, deserves respectful attention" (Sunstein 2016a: 117-118). Nevertheless, and moving to the second part of the statement, it is clear that "there are good reasons why public acceptability should not be the sole or determining condition for going forward with behaviour change" (Dolan et al. 2010: 63). Such data cannot resolve ethical questions because: (i) public responses may be "confused, insufficiently considered, or wrong" (Sunstein 2016a: 116); (ii) there are deontological constraints that apply, regardless of public opinion; and/or (iii) governments may, in some instances, need "to take a lead on issues despite public opposition", and that this sometimes actually leads to a shift in public attitudes in response to the introduction of the policy (Dolan et al. 2010: 63).

I do not have any novel claims about this methodological issue. Rather, I do wish to highlight that the existing data on public attitudes do not pose any challenge to the transformative nudge model, since the data are broadly aligned with it. If anything, they provide the transformative nudge model with additional validation, especially relative to the standard nudge model as it is represented by Sunstein's arguments and the BIT's practical design and implementation of nudge policies. Across the three main existing studies of public attitudes, the basic finding is that "strong majorities were inclined to prefer nudges that appeal to people's deliberative capacities" (Sunstein 2016a: 141). This type of nudging was referred to as "System 2" (versus "System 1") nudges in both Jung and Mellers (2016) and Sunstein (2016b), and as "overt" (versus "covert") nudges in Felsen et al. (2013) – though both sets of terms are defined in the same way. This evidence shows that the nudge-to-deliberate principle is, as a matter of empirical fact, supported.[165]

---

[164] Dolan et al. argue in terms of *legitimacy*: "Government legitimacy rests on the fact it represents and serves the people, and thus it is vital that their views are taken into account when considering any attempt to influence their behaviour" (2010: 63). Sunstein makes a similar claim with reference to *self-government*: Given the "commitment to democratic self-government…whenever public officials are uncertain about whether an approach is desirable, it would be reasonable, in the name of self-government, for them to give consideration to the views of members of the public" (2016a: 118). Others might also make reference to the *value of political autonomy to a flourishing life*: "Just as there is a value to individuals having control over their own lives and choices, independently of the quality of their decisions, so there may be a value to people exercising control over their shared environment and the rules […] that apply to them collectively" (Brighouse et al. 2016: 17).

[165] Although more nuanced experiments that might test for public support for the nudge-to-motivate principle have not yet been designed, some related research into the concept of "pre-authorization"

Third, there is the issue of about nudging's superficiality as a way of responding to social problems. The strong form of this challenge is that the use of nudges, even transformative nudges, is an excuse for not acting more decisively on a particular issue. This worry has two parts: first, these interventions can act only on the effects, rather than on the causes, of certain social problems; and, second, their use provides the government with essentially "a fig leaf for inaction" (Halpern 2015a: 315). In other words, nudges may allow governments to appear as if they are tackling a problem without taking the (often more politically costly) action needed to deal effectively with the problem and its causes.

Research into the relationship between social order in modern societies and people's moral behaviour highlights an illustrative example. This has found that anxiety in the workplace leads people to "focus narrowly on their own basic needs and self-interest", and that this "self-protective mode" can cause them to be "less mindful of principles that guide ethical and moral reasoning – and make them rationalize their own actions as acceptable" (Kouchaki and Desai 2015: 360). Assuming that the research has found a robust link, this raises the question of whether moral nudging is problematically superficial: instead of dealing with some of the underlying issues relating to why people are living their lives in this self-protective mode, and therefore why they are more likely to fail to reason and act in accordance with moral norms, transformative nudges appear only to be able to act on (some of) the surface-level effects of the deeper problem.

A question that follows from this is what proportion of its (limited) resources should governments be giving over to developing such policies, relative to the resources they devote to seeking to deal with the more complex and costly underlying problems causing, in this case, anxiety? Given the pervasiveness of anxiety in modern societies, and its relationship with rising levels of uncertainty and ever-more precarious organisational corporate structures (Standing 2014), citizens might have a legitimate complaint if their government's response to this social ill is to nudge them (even in permissible ways) rather than to address the root causes of these behaviours by, for example, regulating at least the most anxiety-inducing working conditions (e.g., zero-hour contracts). Governments may, in turn, respond by claiming that, in the case of moral nudging at least, it has a duty to act in *some* way and that, in lieu of bringing about

(Niker et al. 2016) suggests that people's attitudes towards different kinds of System 1 nudges – i.e., those that offer motivational scaffolding for behaviours we believe we have reasons to perform, versus those that seek to change our automatic behavioural responses regardless of our beliefs about reasons for action – might also align with the normative analysis presented here.

more politically fraught institutional changes (which can take a long time, even when political will exists), PCA represents one way of addressing a problem.[166]

Even taking into consideration this possible governmental response, it is easy to feel the force of this issue, and rightly so. Yet two related points temper it as a criticism of my account of permissible nudging. The first is again to restate the scope of the ambitions of the thesis and of the transformative nudge model, which has been to answer the question of which forms of public nudging are, in principle, morally permissible. In this inquiry, all of the other policy levers remain in the policymaker's toolkit to be used when necessary and, if the arguments presented in this thesis were extended to these, there would evidently be a huge number of policies that democratic governments should consider pursuing in order to enable citizens to live well – including, no doubt, labour reforms for the sake of anxiety-reduction. Of course, these issues lie beyond the scope of the thesis.

The second point is that nothing in my argument has indicated that nudge policies should be used as the sole or the primary means of addressing social and political problems. By arguing that some forms of PCA are morally permissible, what has been shown is that governments *can* use this form of intervention, wherever appropriate. This appropriateness might come about because other policy levers are unable to address the problem at hand (such as in the case of implicit biases), or sometimes because nudges are more efficacious due to certain properties of the case in hand (note: this does not mean more cost-effective). But they will often be used as a low-cost complement to other, longer-term and more costly forms of intervention. Policy problems are complex, and so policy responses often need to be multi-pronged. Permissible nudges may provide a helpful supplement in such cases, but often cannot, and therefore should not, be offered as the only or often even the main intervention.[167]

A fourth issue is comparative in nature, since it attends to a problem that arises due to how some in the political community are treated in comparison with how others are treated. The particular concern relates to the possibility that some cognitive biases may be worsened, or even caused, by particular social conditions. There are different ways of specifying this kind of concern empirically, but I will do so by reference to

---

[166] This is, in fact, the kind of response given by Halpern. He acknowledges this "efficacy challenge", which he believes comes from "the political left", but responds with the "pragmatist" view that "we should do whatever works, particularly if it has minimal costs" (2015a: 315-317).

[167] This point was clearly made in the House of Lords Select Committee report *Behaviour Change*, the central finding of which is that "non-regulatory measures used in isolation, including 'nudges', are less likely to be effective. Effective policies often use a range of interventions" (2011: 5).

evidence from developmental psychology. As we have seen, the cognitive biases and motivational shortfalls that nudges are designed to overcome tend to be the result of the "uneasy interaction" between two kinds of cognitive processes (Kahneman 2011: 415). Is it the case that some people are, due to their upbringing for instance, more likely to be prone to these decisional and motivational biases? According to Matt Grist, research into how our brains develop in childhood suggests just this. He writes:

> "One crucial implication of brain plasticity is that cognitive and emotional brain processes in children, although innate, do not properly develop without the right kind of nurturing and learning. Children who do not receive the latter are not likely to develop a well-functioning balance between their controlled and automatic brains. If this happens they may be unable to exercise self-control or take into account long-term consequences, which in turn affects their ability to learn [and their capacity to respond to reasons]. Their decision-making and behaviour may also be hijacked by fearful and aggressive responses emanating from their automatic brains. Finally, they may not properly develop emotional capabilities such as empathy for others." (Grist 2009: 43-44)

The issues raised by this quotation are complex and wide-ranging. For our current purposes, however, this brief insight into developmental psychological research raises the question of whether public nudges may, in reality, often operate on certain groups within the population more than others, and whether one relevant factor in this relates to whether or not one was fortunate enough to receive the "right kind of nurturing and learning" in one's early years. This raises further complex questions about the distributional dynamics of nudging across the population. If the transformative nudge model is found to act on certain people more than others, would this *de facto* targeting necessarily cause problems for the permissibility of this account? In order to assess these questions in our all things considered judgements, we would need to know more about the distributional dimensions of PCA. There is currently very little (if any) work, either empirical or normative, on this issue. Hence, rather than make any tentative comments here, I raise this issue in order to highlight an important area for future research – which, in turn, leads us to the chapter's final section.

§6.4: Future Directions

Of necessity, this thesis has focused its attentions on a particular (and currently the most prominent) category of choice architecture – public nudging – and the related set of questions about its permissible use as a means of intentionally influencing people's

behaviour. This final section moves beyond these limitations by briefly sketch some further questions and issues relating to the political implications of the psychological research into human behaviour and decision-making. These issues broadly fall into one of two categories – possible extensions of the analysis of public nudging (§6.4.1) and ethico-political questions that arise when we turn our attention to the use of nudges outside of the public realm (§6.4.2). These represent two ways in which my current theory of the political morality of nudging could be extended. Even though this theory provides many of the resources required for carrying out these extensions, further distinctive questions need to be addressed and the results of these inquiries integrated into the existing account in order to provide full answers to these important extensions. This section can, therefore, be understood as setting out an agenda for future research.

§6.4.1: Further Issues Relating to Public Nudging

Within the first of these categories, there are (at least) two additional issues that would be worth exploring. These relate to the limitations on the thesis's scope concerning the persons whom the nudges have been assumed to be acting upon. This thesis has accounted for a central range of cases, namely, adult citizens assumed to be above a minimum threshold of autonomy. Further research could explore how the permissibility conditions for nudging change when applied to cases outside of this central range, such as those directed towards young children and adolescents, on one side of this range, and previously-autonomous but currently non-autonomous adults such as those with dementia, on the other. The key consideration in these cases is to assess whether there is anything special about children or the demented that narrows or widens the scope of permissible nudging.

With respect to children, there is a widely held belief that the child's lack of autonomy justifies widening the scope of permissible interference beyond that which is permissible in the case of adults and that this grounds, for instance, the compulsory education systems for children that exist in many societies. There are important and ongoing debates about how precisely this more extensive interference should be discharged (e.g., debates about the content of moral and civic education). For our purposes the main question is, with more traditional modes of influence available, what role, if any, is there for nudges within the education system? The fact that the scope of permissible interference is wider in the case of children does not *necessarily* extend to any widening in the scope of permissible *nudging* of children. It might do: for instance, System 1 moral nudges might be considered permissible forms of non-intellectual

character habituation, at least in the case of young children, since they might play an early role in shaping citizens' moral and civic characters.[168] But, if so, arguments would have to be offered to this end, and no such arguments have been provided here. It is also plausible that, given the other options on the table, any such extension of the scope of permissible nudging would, even if permissible in principle, be overridden by the possibility of using more effective or more morally desirable methods. Gigerenzer's suggestions for educational reforms might be one such option. In any case, it is clear that more work needs to be done, first, to clarify precisely what the morally relevant differences are between the two cases and, second, to examine how the theory of the political morality of nudging developed in these pages (or an alternative theory) would apply in the case of children.[169]

With respect to the demented, it seems plausible that there might be several ways in which extreme forms of choice architecture might be used as part of their publicly-funded medical treatment. Take the case of a progressive care home for those with severe dementia in the Netherlands as an example. 'Hogewey' is a gated model village, complete with town square, post office, theatre, hair salon, café-restaurant and supermarket – as well as cameras monitoring residents around the clock, and well-trained staff working incognito, holding a myriad of occupations such as post-office clerks and supermarket cashiers. Every detail of this "fake reality" has been meticulously designed in the hope that the residents can experience life as close to 'normal' as possible.[170] It has been shown that this all-day reminiscence therapy reduces the amount

---

[168] For what is, to the best of my knowledge, the only paper in the literature at the current time addressing the relationship between nudges and habituation, see Niker (forthcoming). Although the socio-ecological account of critical habituation that it develops relates to adults only, see the final section for a brief discussion of its potential implications for traditional educational contexts and, by extension, children. For more on habituation, see Burnyeat (1980), Sherman (1989), and Kristjánsson (2006).

[169] With regards to the first point, see Fowler's 'Perfectionism for Children, Anti-Perfectionism for Adults' (2014) as an example of the view that different criteria apply. Furthermore, this task might include assessing whether certain kinds of nudges negatively impact on "childhood goods", that is, goods that are available *only* in childhood (see Brighouse et al. (2016: 15-17) for more on accounting for these kinds of goods in policymaking). With respect to the second point, the work of Clayton and Moles (forthcoming) and MacMullen (2013) will be instructive – particularly with regards to the case of moral nudging. Also, for real-world examples relating to the case of children, see 'Nudging for Kids': an online platform for creating "creative visual solutions for parents, schools, communities, and organizations to build a positive environment for children and students" (http://www.nudgingforkids.com/).

[170] Hogewey's critics have drawn parallels with the deception depicted in the social-science-fiction film, *The Truman Show*; but many Alzheimer's experts have praised it for being the first to adjust 'our' reality to allow for those with dementia. Taking inspiration from Hogewey, care homes in the UK – such as Grove Care in Winterbourne, Bristol – have developed similar (non-immersive) features; in Grove Care's case, 'Memory Lane': a recreation of a 1950s high street, including a Post Office, pub, bus stop, phone box, and shop windows full of memorabilia. For more on this, see my blog post for *Justice Everywhere* (26th January 2015): http://justice-everywhere.org/old-blog/truman-care-for-dementia.

of medication that residents need to take, and is said to improve their lives in numerous other ways.

There is, of course, the question about whether such forms of influence, given their relationship with medical treatment, can or should be called choice architecture or nudges. In my view, they share the two main features of nudges, since they are neuroscientifically-informed and design-based, though I accept with respect to the former that there is a difference between the evidence relating to cognitive biases and heuristics, on the one hand, and that relating to severe memory loss, on the other. But, given the license granted by a discussion of future research, it is interesting to note that the case of the demented poses an interesting puzzle for my general theory, since this theory rests on an account of living well that posits a particular relation between autonomy and human flourishing. Hence, any application or adaptation of my theory to cases of this sort would require engaging with a distinctive set of normative questions concerning, for instance, what it means to live well as a non-autonomous agent and what duties we have to respect the continuity of life, in addition to caring for the demented person's experiential interests, if any.[171]

This research would speak directly on one of the UK's policy priorities, namely, the "national crisis" in dementia care.[172] Interestingly, the UK government has recently launched a national dementia strategy which they have called *Living Well with Dementia*. There is, however, no specification of what "living well with dementia" means in the policy papers (Department of Health 2009). Hence, returning to the points about the relationship between political theory and policy, this strand of research could link important work in philosophy on this issue with the current development of this policy strategy, and examine the ethical dimensions of the possibility of bringing design-based treatments, such as all-day reminiscence therapy, into mainstream use.

§6.4.2: Shifting Attention to Private Nudging

A second category for future research is the extension of the theory of political morality of nudging beyond the confines of governments and into the private realm. The psychological evidence about the relationship between social organisation and individual agency also raises questions about how governments might permissibly regulate public environments in light of the extensive *private nudging* that citizens' experience, as well as

[171] For the philosophical debate on these issues, see, in particular, Dworkin (1993), Shiffrin (2004), and Jaworska (1999). For a very good overview (and extension) of these arguments, see McKerlie (2012).
[172] For former Prime Minister David Cameron's reference to dementia care as a national crisis, see e.g.: http://www.bbc.co.uk/news/health-17507678.

questions about the responsibilities of the corporate actors themselves to develop ethical best practice guidelines with respect to their ongoing and future development and use of influencing techniques.

Innovations in ways of influencing behaviour are driven by the private sector: by the time nudging was being discussed as a novel means of intervention in the public sector, these kinds of influences had already been in use by advertising and marketing agencies for over half a century (Packard 2007 [1957]). New and emerging techniques for private nudging tend to be highly personalised forms of nonargumentative influence, based on algorithms that work using big data and deep learning ('artificial intelligence').[173] These techniques are making possible manipulations that "have no precedent in human history and that are currently well beyond the scope of existing regulations and laws" (Epstein 2016). They have been made possible by recent developments in technology and society, including: (i) the Internet becoming part of the backdrop of human experience, with devices becoming more persistent companions through life; (ii) rapid advances in measurement and analytics, which are enabling us more quickly to optimise technologies to reach greater levels of persuasiveness (or "behaviour design"); and (iii) the explosion of information and media, which some have referred to as causing "a crisis of distraction" (Williams 2013). Another factor in the development of so-called "persuasive technology" (Fogg 2003) has been the application of knowledge of psychological biases to technology design, and it is this factor, in particular, that brings these persuasive technologies into the debate over nudging.[174]

Several complex and highly pressing issues are raised by private nudging; I am able to touch on only a few of these here. First, there is a conceptual question about how and to what extent these techniques and technologies could be thought of as acting *persuasively*, if at all (Verbeek 2009). As discussed in the previous chapter, persuasion is generally understood – within political theory, at least – as seeking to change a person's behaviour by changing their beliefs or desires through the giving of reasons; so, to what extent is "persuasive technologies" a misnomer for these practices? Or, to what extent can these private nudges be understood via the concept of public-ecological persuasion, introduced in this thesis? Answering these questions requires carrying out careful

[173] For some examples and discussion of these new advertising techniques such as "digital insertion", see a post I wrote on the *Justice Everywhere* blog: http://justice-everywhere.org/economics/freedom-versus-free-stuff (19th November 2015).

[174] This field of research was originally called "captology", a term derived from the acronym 'Computers As Persuasive Technologies' (Fogg 2003). Since then, the practical phenomenon that it examines has exploded: each day, new computing products – be they mobile apps, websites, video games, etc. – are being designed to change people's behaviour and beliefs.

taxonomical and conceptual work in order to assess several of these new and emerging techniques and to categorise them variously as persuasive, manipulative, or even coercive.

Second, it is important that liberal political theory engages seriously and systematically with the regulatory and moral issues raised by these technological developments and the powers they have conferred on private nudging. The debate over nudging (this thesis included) can be said to have suffered, thus far at least, from a "libertarian bias" since it has focused on the dangers created by governmental action without considering to the same extent "whether private action might present comparable dangers or whether the government might help overcome the dangers created by private action" (Strauss 1991: 361). Since it is clear that private nudging is at least as dangerous as public nudging, arguably the main question that needs addressing is whether, and under what conditions, governments might be permitted to constrain or regulate such private influencing activities (see, e.g., Moore 2017).[175]

Addressing this question would require engaging with a host of normative questions relating to, for instance, the freedom of expression of private actors (and, in the US, its protection by the First Amendment), the privacy claims of citizens, and issues around informed consent.[176] It would also require further consideration of the ethics of "counter-manipulation", that is, when certain manipulative techniques are used as a means of countering manipulation (Wilkinson 2016). What difference does it make to the ethics of manipulation when it is *counter*-manipulation? In this case, might otherwise impermissible public nudges be used permissibly for the sake of countering the manipulative efforts of private nudges?[177] Sunstein does offer a brief discussion of what he calls "regulating manipulation", which is rather pessimistic on account revolving around the severity of the free speech barriers (2016a: 111; see also Strauss 1991). A fuller discussion might look to apply to private nudging the theory of the

[175] Indeed, "Denmark has recently announced that it would be creating a new diplomatic posting – a "digital ambassador" – to deal not with states and international organizations, but with giant technology companies like Facebook and Google" (Taylor 2017). For an interview with Anders Sandersen, the Danish Foreign Minister, detailing why such a position is deemed necessary and what it might look like in reality, see Taylor (2017).

[176] With respect to the last of these, consent, it is important to note that Facebook defended the huge-scale emotional priming experiment, which manipulated over 700,000 users' news feeds in order to see how this might affect the positivity or negativity of their own posts (Kramer et al. 2014), "on the grounds that its users' consent to their terms of service was sufficient to imply consent to such experiments" (Shaw 2017).

[177] Returning to our discussion of nudging children, too, there is a further question to those outlined in §6.4.1 relating to *counter-nudging* for the sake of protecting them (and their nascent autonomy) from child advertising.

political morality of nudging developed in these pages (or to apply an alternative account). This task would require carrying out the conceptual analysis and addressing the distinctive normative questions outlined above, before factoring this new work into those considerations about well-being, autonomy, and moral reasoning – and the appropriate role of the political community in protecting and promoting these values – already captured by the theory.[178]

Third, there is now a pressing public debate about the ethical guidelines that should apply to the ongoing and future development of "persuasive technologies". In recent months, the world's most (persuasively) powerful global corporations – Google, Facebook, Amazon, IBM, Microsoft, and Apple (which joined later) – have come together to form the "Partnership on AI" in order to set societal and ethical best practice guidelines for artificial intelligence research and development.[179] This is a welcome advance in light of the vast power held and exerted by these corporations as "Attention Brokers" (Wu 2017) within what is now often referred to as the "attention economy" (Davenport and Beck 2001; Terranova 2012).[180] Future research can, alongside this corporate partnership, consider the issue of what sorts of (new) social responsibilities accrue to those corporations acting as attention brokers, and how any such responsibilities should be understood (for example, are they *collective* responsibilities and, if so, by which mechanisms might we be able to hold these corporations to account?).

Such questions are all the more pressing in light of recent events and the move into what many have described as a "post-truth world" of "fake news" and "alternative facts" (see, e.g., Levy 2017). Given Facebook's role as a news source and its ability to manipulate what we see and subsequently how we feel (Kramer et al. 2014), and given the power in Google's hands to decide which webpages to include in search results and how to rank them (Epstein 2016), there are important questions about whether there are any circumstances under which choice architecture might permissibly be used to nudge citizens towards particular evidence-based news sources or away from categorically fake news sources.

---

[178] I view this extension, which I plan to carry out as my postdoctoral research project, as being the second part of a two-part broader research project, entitled "Living Well, by Design" – of which this thesis is the first part.

[179] For more information, see: https://www.partnershiponai.org.

[180] "Attention brokers" are firms that attract and resell attention (or, more precisely, the data produced when one's attention is given to, e.g., Facebook) to advertisers.

§6.5: Conclusion

This chapter has sought to present in its essential form the central thrust of the normative analysis offered in the preceding chapters and to engage with more practical concerns and issues relating to its impact on the way in which behavioural policies are designed, especially within the UK context. On the first, it has provided an account of the agency-enhancing potential of PCA and has set out two ethical principles that follow from this so-called transformative nudge model. On the second, it has engaged with current policymaking practices, and has presented an ethically sensitive version of the BIT's existing policy development framework, which represents the first attempt, to the best of my knowledge, to integrate ethical concerns into current policy practices. It has also outlined and reviewed some of the limitations of this means of intervention (generally speaking), which would also apply to my transformative nudge model and to its successful and all things considered permissible implementation in practice. Finally, the chapter has discussed some of the limitations of the argument presented in these pages, which principally pick out ways in which it could be extended to include further issues that relate to both public and private nudging.

7. Conclusion: Living Well by Design

We are living in an age of the behavioural sciences. Our updated understanding of human agency is increasingly being used as the basis for designing environments that direct behaviour in particular ways, often without us being aware. These developments call for serious philosophical engagement to clarify the conceptual and normative terrain and to assess the permissibility of such influences; for, although behavioural policy holds out some promise for helping us to live well, it also comes with the risk of significant ethical and moral costs. This thesis has responded to this call. Focusing on the governmental use of these design-based behaviour change strategies, it has defended a complete account of the political morality of public nudging, which takes advantage of the promise of PCA while clarifying and, as a result, managing to avoid its normative costs. It is an account of *living well by design*, as this relates to governmental behavioural policymaking.

Rather than rehearse my arguments in favour of this account, I wish to conclude by briefly issuing a few concluding remarks. First, this thesis has investigated one particular instance of how the recent discoveries about human nature unsettle orthodox positions in political theory. There is much more to be done in considering the potential implications of these new models of human agency. All systems of normative theory and practical ethics make certain presuppositions about human agents; but, if people are shown to lack these assumed capacities in a systematic way, as has been shown by the evidence, then "the many practical injunctions that follow from ethical discourse come to seem misguided" (Mele and Shepherd 2013: 62). The account of permissible public influence offered in this thesis is distinct from traditional autonomy-based accounts in the literature, because it has sought to take seriously the internal threat to autonomy from the psychological findings relating to our automatic and context-sensitive mental processes. This has had important implications for the normative analysis. I hope that other political theorists will engage in a similar way with the empirical research in order to explore in more detail its various (applied) theoretical implications.

Second, the purpose of this thesis has not been to solve a philosophical puzzle. Although it has engaged with a full range of conceptual and normative issues, often at a high level of philosophical abstraction, the thesis started with, and has sought to remain disciplined by, a subject of practical political importance. In so doing, it has been able to deliver a bespoke theory of political morality for public nudging, and it has used this as

the basis for exploring what an ethically-sensitive behavioural policy framework might look like in practice. This was done by integrating the relevant normative considerations into the current practices and framework of the BIT, to show that this would not need to involve a wholesale revision of its existing intervention design process. There are, no doubt, many further practical issues that need to be taken into account. Nevertheless, the thesis's efforts to present its argument in an accessible way marks a genuine attempt to start a more productive, two-way discussion between political theorists and policy practitioners about the inclusion of ethical considerations into the design framework for nudging – something that is increasingly urgent in light of the recent vast expansion of behavioural policy.

Bibliography

Adams, R.M. (2006), *A Theory of Virtue: Excellence in Being for the Good*, Oxford: Oxford University Press.

Adolphs, R., and Damasio, A. (2001), 'The Interaction of Affect and Cognition: A Neurobiological Perspective', in J.P. Forgas (ed.), *Handbook of Affect and Social Cognition*, Mahwah, NJ.: Erlbaum, pp. 27-49.

Ahlstrom-Vij, K. (2013), *Epistemic Paternalism: A Defence*, Basingstoke: Palgrave Macmillan.

Anderson, J. (2010), 'Review: Nudge: Improving Decisions about Health, Wealth, and Happiness', *Economics and Philosophy* 26(3), pp. 369-406.

Anscombe, G.E. (1958), 'Modern Moral Philosophy', *Philosophy* 33(124), pp. 1-19.

Ariely, D. (2009), *Predictably Irrational: The Hidden Forces that Shape Our Decisions*, London: Harper.

Aristotle (2009), *The Nicomachean Ethics* (trans. David Ross; revised by Lesley Brown), Oxford: Oxford University Press.

Arneson, R. (2016), 'Does Fairness Require a Multidimensional Approach?', in M.D. Adler and M. Fleurbaey (eds.), *The Oxford Handbook of Well-Being and Public Policy*, Oxford: Oxford University Press, pp. 588-614.

Arpaly, N. (2002), *Unprincipled Virtue: An Inquiry into Moral Agency*, Oxford: Oxford University Press.

Bagnoli, C. (ed.) (2011), *Morality and the Emotions*, Oxford: Oxford University Press.

Bandura, A., Barbaranelli, C., Caprara, G., and Pastorelli, C. (1996), 'Mechanisms of Moral Disengagement in the Exercise of Moral Agency', *Journal of Personality and Social Psychology* 71, pp. 364-374.

Barber, B. (1971), *Superman and Common Men: Freedom, Anarchy, and the Revolution*, New York: Praeger.

Bargh, J.A., Chen, M., and Burrows, L. (1996), 'Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action', *Journal of Personality and Social Psychology* 71(2), pp. 230-244.

Bargh, J.A., and Chartrand, T.L. (1999), 'The Unbearable Automaticity of Being', *American Psychologist* 54(7), pp. 462-479.

Baron, J. (2016), 'A Welfarist Approach to Manipulation', *Journal of Marketing Behavior* 1(3), pp. 283-291.

Baron, M. (2014), 'The *Mens Rea* and Moral Status of Manipulation', in C. Coons and M. Weber (eds.), *Manipulation: Theory and Practice*, Oxford: Oxford University Press, pp. 98-120.

Bator, F.M. (1958), 'The Anatomy of Market Failure', *Quarterly Journal of Economics* 72(3), pp. 351-379.

Batson, C.D., and Thompson, E.R. (2001), 'Why Don't Moral People Act Morally?', Motivational Considerations', *Current Directions in Psychological Science* 10(2), pp. 54-57.

Beck, U. (1992), *Risk Society: Towards a New Modernity*, London: Sage Publications.

Beckman, L. (2001), *The Liberal State and the Politics of Virtue*, London: Transaction Publishers.

Begon, J. (2016), 'Recent Work: Paternalism', *Analysis* 76(3), pp. 355-373.

Behavioural Insights Team (2015), 'Update Report 2013-2015', London: BIT. Available online at: http://www.behaviouralinsights.co.uk/publications/the-behavioural-insights-team-update-report-2013-2015.

--------- (2016), 'Update Report 2015-2016', London: BIT. Available online at: http://www.behaviouralinsights.co.uk/publications/the-behavioural-insights-teams-update-report-2015-16.

Ben-Porath, S. (2010), *Tough Choices: Structured Paternalism and the Landscape of Choice*, Princeton: Princeton University Press.

Benforado, A. (2015), *Unfair: The New Science of Criminal Justice*, New York: Crown Publishing.

Berker, S. (2009), 'The Normative Insignificance of Neuroscience', *Philosophy & Public Affairs* 37(4), pp. 293-329.

Berlin, I. (2002 [1969]), 'Two Concepts of Liberty', in I. Berlin, *Four Essays on Liberty*, Oxford: Oxford University Press, pp. 118-172.

Binder, M., and Lades, L.K. (2015), 'Autonomy-Enhancing Paternalism', *Kyklos* 68(1), pp. 3-27.

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

--------- (2013), 'Choice Architecture: A Mechanism for Improving Decisions While Preserving Liberty?', in C. Coons and M. Weber (eds.), *Paternalism: Theory and Practice*, Cambridge: Cambridge University Press, pp. 178-196.

--------- (2014), 'A Framework for Assessing the Moral Status of "Manipulation"', in C. Coons and M. Weber (eds.), *Manipulation: Theory and Practice*, Oxford: Oxford University Press, pp. 121-134.

--------- (2016), 'Biases and Heuristics in Decision-Making and their Impact on Autonomy,' *The American Journal of Bioethics* 16(5), pp. 5-15.

Blumenthal-Barby, J.S., and Krieger, H. (2015), 'Cognitive Biases and Heuristics in Decision-Making: A Critical Review Using a Systematic Search Strategy', *Medical Decision Making* 35(4), pp. 539-557.

Bovens, L. (2009), 'The Ethics of *Nudge*', in T. Grüne-Yanoff and S.O. Hansson (eds.), *Preference Change: Approaches from Philosophy, Economics and Psychology*, Dordrecht: Springer, pp. 207-220.

Brady, M.S. (2013), *Emotional Insight: The Epistemic Role of Emotional Experience*, Oxford: Oxford University Press.

Bratman, M. (2005), 'Planning Agency, Autonomous Agency', in J.S. Taylor (ed.), *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, Cambridge: Cambridge University Press, pp. 33-57.

Brighouse, H., and Swift, A. (2014), *Family Values: The Ethics of Parent-Child Relationships*, Oxford: Princeton University Press.

Brighouse, H., Ladd, H.F., Loeb, S., and Swift, A. (2016), 'Educational Goods and Values: A Framework for Decision Makers', *Theory and Research in Education* 14(1), pp. 3-25.

Broder, J.M. (2012), 'Powerful Shaper of US Rules Quits, With Critics in Wake', *The New York Times* (3rd August): http://www.nytimes.com/2012/08/04/science/earth/ cass-sunstein-to-leave-top-regulatory-post.html.

Brown, P. (2012), 'A Nudge in the Right Direction? Towards a Sociological Engagement with Libertarian Paternalism', *Social Policy and Society* 11(3), pp. 305-317.

Brownlee, K. (2015), 'What's Virtuous about the Law?', *Legal Theory* 21(1), pp. 1-17.

Brownlee, K., and Child, R. (2012), 'Can the Law Help Us to be Moral?', Warwick School of Law Research Paper (Number 2012/17).

Brownstein, M. (2015), 'Implicit Bias', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.

--------- (2016), 'Attributionism and Moral Responsibility for Implicit Bias', *Review of Philosophy and Psychology* 7(4), pp. 765-786.

Brownstein, M., and Saul, J. (eds.) (2016a), *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, Oxford: Oxford University Press.

--------- (eds.) (2016a), *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*, Oxford: Oxford University Press.

Bubb, R., and Pildes, R. (2014), 'How Behavioral Economics Trims its Sails and Why', *Harvard Law Review* 127, pp. 1593-1678.

Bublitz, C. (2016), 'Moral Enhancement and Mental Freedom', *Journal of Applied Philosophy* 33(1), pp. 88-106.

Burnell, P., and Reeve, A. (1984), 'Persuasion as a Political Concept', *British Journal of Political Science* 14(4), pp. 393-410.

Burnyeat, Myles (1980), 'Aristotle on Learning How to be Good', in A.O. Rorty (ed.), *Essays on Aristotle's Ethics*, Berkeley: University of California Press, pp. 69-92.

Buss, S. (2013), 'Personal Autonomy', in E.N. Zalta (ed.)*, The Stanford Encyclopedia of Philosophy*.

Cabinet Office (2010), 'Applying Behavioural Insight to Health', London: Cabinet Office. Available at: https://www.gov.uk/government/publications/applying-behavioural-insight-to-health-behavioural-insights-team-paper.

--------- (2012), 'Tackling Fraud and Error in Government', London: Cabinet Office. Available at: https://www.gov.uk/government/publications/tackling-fraud-and-error-in-government-a-report-of-the-fraud-error-and-debt-taskforce.

--------- (2013), 'Applying Behavioural Insights to Charitable Giving', London: Cabinet Office. Available at: http://www.behaviouralinsights.co.uk/publications/applying-behavioural-insights-to-charitable-giving.

Callan, E.K. (2015), 'Debate: Liberal Virtues and Civic Education', *The Journal of Political Philosophy* 23(4), pp. 491-500.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003), 'Regulation for Conservatives: Behavioral Economics and the Case for "Asymmetric Paternalism"', *University of Pennsylvania Law Review* 151(3), pp. 1211-1254.

Caney, S. (1995), 'Anti-Perfectionism and Rawlsian Liberalism', *Political Studies* 43, pp. 248-264.

Carter, I. (2011), 'Respect and the Basis of Equality', *Ethics* 121, pp. 538-571.

Cartwright, N. (2010), 'What are Randomised Controlled Trials Good For?', *Philosophical Studies* 147(1), pp. 59-70.

Chaiken, S., and Trope, Y. (eds.) (1999), *Dual-Process Theories in Social Psychology*, London: The Guilford Press.

Chan, J. (2000), 'Legitimacy, Unanimity, and Perfectionism', *Philosophy & Public Affairs* 29(1), pp. 5-42.

Chittaranjan, G., Blom, J., and Gatica-Perez, D. (2013), 'Mining Large-Scale Smartphone Data for Personality Studies', *Personality and Ubiquitous Computing* 17(3), pp. 433-450.

Christman, J. (2009a), 'Autonomy in Moral and Political Philosophy', in E.N. Zalta (ed.)*, The Stanford Encyclopedia of Philosophy*.

--------- (2009b), *The Politics of Persons: Individual Autonomy and Socio-historical Selves*, Cambridge: Cambridge University Press.

--------- (2017), 'Anti-Perfectionism and Autonomy in an Imperfect World: Comments on Joseph Raz's *The Morality of Freedom* 30 Years On', *Moral Philosophy and Politics* (doi: 10.1515/mop-2016-0037).

Christman, J., and Anderson, J. (2005), 'Introduction', in J. Christman and J. Anderson (eds.), *Autonomy and the Challenges to Liberalism: New Essays*, New York: Cambridge University Press, pp. 1-23.

Cialdini, R.B. (2003), 'Crafting Normative Messages To Protect The Environment', *Current Directions in Psychological Science* 12(4), pp. 105-109.

Cialdini, R.B., Demaine, L.J., Barrett, D.W., Rhoads, K., and Winter, P.L. (2006), 'Managing Social Norms for Persuasive Impact', *Social Influence* 1(1), pp. 3-15.

Clarke, S. (2006), 'State Paternalism, Neutrality and Perfectionism', *The Journal of Political Philosophy* 14(1), pp. 111-121.

Clayton, M. (1993), 'White on Autonomy, Neutrality and Well-Being', *Journal of Philosophy of Education* 27(1), pp. 101-113.

--------- (2006), *Justice and Legitimacy in Upbringing*, Oxford: Oxford University Press.

Clayton, M., and Moles, A. (forthcoming), 'Neurointerventions, Morality, and Children', in D. Birks and T. Douglas (eds.), *Treatment for Crime*, Oxford: Oxford University Press.

Colburn, B. (2010), *Autonomy and Liberalism*, Abingdon: Routledge.

--------- (2011), 'Autonomy and Adaptive Preferences', *Utilitas* 23(1), pp. 52-71.

Conly, S. (2013), *Against Autonomy: Justifying Coercive Paternalism*, Cambridge: Cambridge University Press.

Connelly, J. (2014), 'Idealism, the Common Good, and Environmental Virtues', in T. Brooks (ed.), *Ethical Citizenship: British Idealism and the Politics of Recognition*, Basingstoke: Palgrave Macmillan, pp. 224-244.

Coons, C., and Weber, M. (eds.) (2013), *Paternalism: Theory and Practice*, Cambridge: Cambridge University Press.

--------- (eds.) (2014), *Manipulation: Theory and Practice*, Oxford: Oxford University Press.

Crisp, R. (2013), 'Well-Being,' in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.

Crisp, R., and Hooker, B. (eds.) (2000), *Well-Being and Morality: Essays in Honour of James Griffin*, Oxford: Clarendon Press.

Crockett, M.J., Clark, L., Tabibnia, G., Lieberman, M.D., and Robbins, T.W. (2008), 'Serotonin Modulates Behavioral Reactions to Unfairness', *Science* 320(5884), p. 1739.

Crockett, M.J., Clark, L., Hauser, M.D., and Robbins, T.W. (2010a), 'Serotonin Selectively Influences Moral Judgement and Behavior through Effects on Harm Aversion', *Proceedings of the National Academy of Science* 107(40), pp. 17433-17448.

--------- (2010b), 'Reply to Harris and Chan: Moral Judgment is More than Rational Deliberation', *Proceedings of the National Academy of Science* 107(50), E184.

Damasio, A. (1994), *Descartes' Error: Emotion, Reason, and the Human Brain*, New York: Putnams Publishing.

Darwall, S. (2006), *The Second-Person Standpoint: Morality, Respect, and Accountability*, Cambridge, MA.: Harvard University Press.

Dasgupta, N. (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, pp. 233-279.

Davenport, T.H., and Beck, J.C. (2001), *The Attention Economy: Understanding the New Currency of Business*, Cambridge, MA.: Harvard Business School Press.

De Dreu, C.K.W., Greer, L.L., Handgraaf, M.J.J., Shalvi, S., Van Kleef, G.A., Baas, M., Ten Velden, F.S., Van Dijk, E., and Feith, S.W.W. (2010), 'The Neuropeptide Oxytocin Regulates Parochial Altruism in Intergroup Conflict Among Humans', *Science* 328, pp.1408-1411.

De Sousa, R. (1997), *The Rationality of Emotion*, Cambridge, MA: The MIT Press.

DeGrazia, D. (2013), 'Moral Enhancement, Freedom, and What we (Should) Value in Moral Behaviour', *Journal of Medical Ethics* 40, pp. 361-368.

Deonna, J.A. (2006), 'Emotion, Perception, and Perspective', *Dialectica* 60(1), pp. 29-46.

Deonna, J.A., and Teroni, F. (2012), *The Emotions: A Philosophical Introduction*, Abingdon: Routledge.

Department of Health (2009), 'Living Well With Dementia: A National Dementia Strategy', London: DH Publications. Available online at: https://www.gov.uk/government/publications/living-well-with-dementia-a-national-dementia-strategy.

Desai, S., and Gino, F. (unpublished), 'Mahatma Gandhi, Email Signatures, and Moral Decisions: The Power of Ethical Nudges', Working Paper. Available at: https://ethics.harvard.edu/ethical-nudges.

Dolan, P., Hallsworth, M., Halpern, D., King, D., and Vlaev, I. (2010), 'MINDSPACE: Influencing Behaviour Through Public Policy', London: Institute for Government. Available at: http://www.behaviouralinsights.co.uk/publications/mindspace.

Doris, J. (2002), *Lack of Character: Personality and Moral Behavior*, Cambridge: Cambridge University Press.

Douglas, T. (2008), 'Moral Enhancement', *Journal of Applied Philosophy* 25(3), pp. 228-245.

--------- (2011), 'Moral Enhancement via Direct Emotion Modulation: A Reply to Harris', *Bioethics* 27(3), pp. 160-168.

--------- (2014), 'Enhancing Moral Conformity and Enhancing Moral Worth', *Neuroethics* 7, pp. 75-91.

Dovidio, J., Kawakami, K., Johnson, C., Johnson, B., and Howard, A. (1997), 'The Nature of Prejudice: Automatic and Controlled Processes', *Journal of Experimental Social Psychology* 33, pp. 510-540.

Duff, R.A. (2001), *Punishment, Communication, and Community*, Oxford: Oxford University Press.

Dworkin, G. (1988), *The Theory and Practice of Autonomy*, Cambridge: Cambridge University Press.

--------- (2005), 'Moral Paternalism', *Law and Philosophy* 24(3), pp. 305-319.

--------- (2017), 'Paternalism', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.

Dworkin, R. (1986), *Law's Empire*, Cambridge, MA.: Harvard University Press.

--------- (1990), 'Foundations of Liberal Equality', in G.B. Peterson (ed.), *The Tanner Lectures on Human Values, Volume 11*, Salt Lake City: University of Utah Press, pp. 1-119.

--------- (1993), *Life's Dominion: An Argument Abortion, Euthanasia, and Individual Freedom*, New York: Alfred A. Knopf.

--------- (1997), 'In Praise of Theory', *Arizona State Law Journal* 29, pp. 353-376.

--------- (2011), *Justice for Hedgehogs*, London: Harvard University Press.

Elster, J. (1985), *Sour Grapes: Studies in the Subversion of Rationality*, Cambridge: Cambridge University Press.

--------- (2000), *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints*, Cambridge: Cambridge University Press.

Engberg-Pedersen, T. (1983), *Aristotle's Theory of Moral Insight*, Oxford: Clarendon Press.

Epstein, R. (2016), 'The New Mind Control', *Aeon* (18th February 2016). Available at: https://aeon.co/essays/how-the-internet-flips-elections-and-alters-our-thoughts.

Evans, J. (2003), 'In Two Minds: Dual-Process Accounts of Reasoning', *Trends in Cognitive Sciences* 7(10), pp. 454-459.

Evans, J., and Stanovich, K.E. (2013), 'Dual Process Theories of Higher Cognition: Advancing the Debate', *Perspectives on Psychological Science* 8(3), pp. 223-241.

Fabre, C. (2006), *Whose Body is it Anyway? Justice and the Integrity of the Person*, Oxford: Clarendon Press.

Faden, R., and Beauchamp, T. (1986), *A History and Theory of Informed Consent*, Oxford: Oxford University Press.

Farrelly, C. (2007), 'Justice in Ideal Theory: A Refutation', *Political Studies* 55, pp. 844-864.

Feinberg, J. (1987), *Harm to Others (Moral Limits of the Criminal Law Volume 1)*, Oxford: Oxford University Press.

--------- (1989), *Harm to Self (Moral Limits of the Criminal Law Volume 3)*, Oxford: Oxford University Press.

Felsen, G., and Reiner, P.B. (2011), 'How the Neuroscience of Decision Making Informs Our Conception of Autonomy', *AJOB Neuroscience* 2(3), pp. 3-14.

Felsen, G., Castelo, N., and Reiner, P.B. (2013), 'Decisional Enhancement and Autonomy: Public Attitudes towards Overt and Covert Nudges', *Judgment and Decision-Making* 8(3), pp. 202-2013.

Fenton, B.D.C. (2014), 'Autonomy, Automaticity, and Attention: Why Empirical Research on Consciousness Matters to Autonomous Agency', PhD thesis. Available at: https://yorkspace.library.yorku.ca/xmlui/handle/10315/29887.

Fischer, J.M. (1999), 'Recent Work on Moral Responsibility', *Ethics* 110, pp. […].

--------- (2012), *Deep Control: Essays on Free Will and Value*, Oxford: Oxford University Press.

Fischer, J.M., and Ravizza, M. (1998), *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

Fogg, B.J. (2003), *Persuasive Technology: Using Computers to Change What We Think and Do*, San Francisco, CA.: Morgan Kaufmann Publishers.

Fowler, T. (2014), 'Perfectionism for Children, Anti-Perfectionism for Adults', *Canadian Journal of Philosophy* 44(3-4), pp. 305-323.

Fowler, T., and Stemplowska, Z. (2015), 'The Asymmetry Objection Rides Again: On the Nature and Significance of Justificatory Disagreement', *Journal of Applied Philosophy* 32, pp. 133-146.

Frankfurt, H.G. (1969), 'Alternative Possibilities and Moral Responsibility', *The Journal of Philosophy* 66(23), pp. 829-839.

--------- (1971), 'Freedom of the Will and the Concept of the Person', *The Journal of Philosophy* 68(1), pp. 5-20.

--------- (1973), 'Coercion and Moral Responsibility', in T. Honderich (ed.), *Essays on Freedom of Action*, London: Routledge and Kegan Paul, pp. 65-86.

Fricker, M. (2007), *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford: Oxford University Press.

Friedman, M., and Friedman, R. (1980), *Free to Choose: A Personal Statement*, London: Harcourt.

Goldstein, D.G., Johnson, E.J., Herrmann, A., and Heitmann, M. (2008), 'Nudging Your Customers towards Better Choices', *Harvard Business Review* (December edition). Available at: https://hbr.org/2008/12/nudge-your-customers-toward-better-choices.

Goodin, R.E. (1982), *Political Theory and Public Policy*, London: The University of Chicago Press.

Gigerenzer, G. (2010), 'Collective Statistical Illiteracy', *Archives of Internal Medicine* 170, pp. 468-469.

--------- (2014), *Risk Savvy: How to Make Good Decisions*, London: Penguin.

--------- (2015), 'On the Supposed Evidence for Libertarian Paternalism', *Review of Philosophy and Psychology* 6(3), pp. 361-383.

Gino, F., and Desai. S. (2011), 'Memory Lane and Morality: How Childhood Memories Promote Prosocial Behavior', *Journal of Personality and Social Psychology* 102(4), pp. 743-758.

Goodwin, T. (2012), 'Why We Should Reject 'Nudge'', *Politics* 32(2), pp. 85-92.

Graham, G. (2015), 'Behaviorism', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.

Griffin, J. (1986), *Well-Being: Its Meaning, Measurement and Moral Importance*, Oxford: Clarendon Press.

--------- (1991), 'Against the Taste Model,' in J. Elster and J.E. Roemer (eds.), *Interpersonal Comparisons of Well-Being*, Cambridge: Cambridge University Press, pp. 45-69.

--------- (1998), *Value Judgement: Improving Our Ethical Beliefs*, Oxford: Oxford University Press.

Grill, K. (2014), 'Expanding the Nudge: Designing Choice Contexts and Choice Contents', *Rationality, Markets and Morals* 5, pp. 139-162.

Grist, M. (2009), 'Changing the Subject: How New Ways of Thinking about Human Behaviour Might Change Politics, Policy and Practice', RSA, London.

Groll, D. (2012), 'Paternalism, Respect, and the Will', *Ethics* 122(4), pp. 692-720.

Groot, B., Sanders, M., Bloomenthal, E., and Rogers, T. (2017), 'I Get by with a Little Help from my Friends: Two Field Experiments on Social Support and Attendance in Further Education Colleges in the UK' (April 2017). Available at: http://38r8om2xjhhl25mw24492dir.wpengine.netdna-cdn.com/wp-content/uploads/2017/04/Study-Supporter-WP_April-2017.pdf.

Grüne-Yanoff, T. (2009), 'Welfare Notions for Soft Paternalism,' *Papers on Economics and Evolution* #0917. Available at: https://www.econstor.eu/handle/10419/32665.

--------- (2012), 'Old Wine in New Casks: Libertarian Paternalism Still Violates Liberal Principles', *Social Choice and Welfare* 38, pp. 635-645.

Grüne-Yanoff, T., and Hertwig, R. (2016), 'Nudge versus Boost: How Coherent are Policy and Theory?', *Minds and Machines* 26(1), pp. 149-183.

Haidt, J. (2006), *The Happiness Hypothesis: Putting Ancient Wisdom and Philosophy to the Test of Modern Science*, London: Arrow Books.

Hallsworth, M., and Sanders, M. (2016), 'Nudge: Recent Developments in Behavioural Science and Public Policy', in F. Spotswood (ed.), *Beyond Behaviour Change: Key*

Issues, Interdisciplinary Approaches and Future Directions, Bristol: Policy Press, pp. 113-133.

Halpern, D. (2015a), *Inside the Nudge Unit: How Small Changes Can Make a Big Difference*, London: WH Allen.

--------- (2015b), 'The Rise of Psychology in Policy: The UK's de facto Council of Psychological Science Advisers', *Perspectives on Psychological Science* 10(6), pp. 768-771.

Halpern, D., Bates, C., Mulgan, G., Aldridge, S., Beales, G., and Heathfield, A. (2004), 'Personal Responsibility and Changing Behaviour: The State of Knowledge and Its Implications for Public Policy', London: The Cabinet Office Strategy Unit. Available online at: http://webarchive.nationalarchives.gov.uk/+/http:/www.cabinetoffice.gov.uk/media/cabinetoffice/strategy/assets/pr2.pdf

Halpern, D., and Sanders, M. (2016), 'Nudging by Government: Progress, Impact, and Lessons Learned', *Behavioral Science & Policy* 2(2), pp. 53-65.

Hand, M. (2008), 'What Should We Teach as Controversial? A Defense of the Epistemic Criterion', *Educational Theory* 58(2), pp. 213-228.

Hanna, J. (2015), 'Libertarian Paternalism, Manipulation, and the Shaping of Preferences', *Social Theory and Practice* 41(4), pp. 618-643.

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

Harman, G. (1999), 'Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error', *Proceedings of the Aristotelian Society* 99, pp. 315-331.

Harris, J. (2011), 'Moral Enhancement and Freedom', *Bioethics* 25(2), pp. 102-111.

--------- (2013a), 'Moral Progress and Moral Enhancement', *Bioethics* 27(5), pp. 285-290.

--------- (2013b), "Ethics is for Bad Guys!' Putting the 'Moral' Back into Moral Enhancement', *Bioethics* 27(3), pp. 169-173.

Harsanyi, J.C. (1977), 'Rule Utilitarianism and Decision Theory', *Erkenntnis* 11(1), pp. 25-53.

--------- (1982), 'Morality and the Theory of Rational Behaviour,' in A. Sen and B. Williams (eds.), *Utilitarianism and Beyond*, Cambridge: Cambridge University Press, pp. 39-62.

Hausman, D.M., and Welch, B. (2010), 'Debate: To Nudge or Not to Nudge,' *The Journal of Political Philosophy* 18(1), pp. 123-136.

Hedlin, S., and Sunstein, C.R. (2016), 'Does Active Choosing Promote Green Energy Use? Experimental Evidence', *Ecology Law Quarterly* 43(1), pp. 107-141.

Hédoin, C. (2016), 'Normative Economics and Paternalism: The Problem with the Preference-Satisfaction Account of Welfare', *Constitutional Political Economy* (published online first: 12th October; doi: 10.1007/s10602-016-9227-5).

Heffernan, M. (2012), *Wilful Blindness: Why We Ignore the Obvious*, London: Simon & Schuster.

Helm, B.W. (2001), *Emotional Reason: Deliberation, Motivation, and the Nature of Value*, Cambridge: Cambridge University Press.

Herstein, O.J. (2012), 'Defending the Right to Do Wrong', *Law and Philosophy* 31, pp. 343-365.

Hertwig, R., and Ryall, M.D. (unpublished), 'Nudge vs. Boost: Agency Dynamics Under 'Libertarian Paternalism' (January 1st, 2016)'. Available at: https://ssrn.com/abstract=2711166.

Holroyd, J. (2012), 'Responsibility for Implicit Biases', *Journal of Social Philosophy* 43, pp. 274-306.

Holroyd, J., and Kelly, D. (2016), 'Implicit Bias, Character, and Control', in A. Masala and J. Webber (eds.), *From Personality to Virtue: Essays on the Philosophy of Character*, Oxford: Oxford University Press, pp. 106-133.

Hosking, K., and Russell, R. (2016), 'Discrimination Law, Equality Law, and Implicit Bias', in M. Brownstein and J. Saul (eds.), *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice and Ethics*, Oxford: Oxford University Press, pp. 254-278.

House of Lords (2011), 'Behaviour Change', Science and Technology Select Committee (Second Report of Session 2010-12), HL Paper 179, London: The Stationery Office Limited. Available at: https://www.publications.parliament.uk/pa/ld201012/ldselect/ldsctech/179/179.pdf.

Howard, J.W. (2017), 'Punishment as Moral Fortification', *Law and Philosophy* 46(1), pp. 45-75.

Huebner, B. (2016), 'Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition', in M. Brownstein and J. Saul (eds.), *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, Oxford: Oxford University Press, pp. 47-79.

Hurley, S. (2011), 'The Public Ecology of Responsibility,' in C. Knight and Z. Stemplowska (eds.), *Responsibility and Distributive Justice*, Oxford: Oxford University Press, pp. 187-217.

Husak, D.N. (1980), 'Paternalism and Autonomy', *Philosophy & Public Affairs* 10(1), pp. 27-46.

Ishmali M'hamdi, H., Hilhorst, M., Steegers, E.A.P., and de Beaufort, I. (2017), 'Nudge Me, Help My Baby: On Other-Regarding Nudges', *Journal of Medical Ethics* (published online first: 25 January 2017, doi: 10.1136/medethics-2016-103656).

Jaworska, A. (1999), 'Respecting the Margins of Agency: Alzheimer's Patients and the Capacity to Value', *Philosophy & Public Affairs* 28(2), pp. 105-138.

Jennings, B., Wertz, F.J., and Morrissey, M.B. (2016), 'Nudging for Health and the Predicament of Agency: The Relational Ecology of Autonomy and Care,' *Journal of Theoretical and Philosophical Psychology* 36(2), pp. 81-99.

John, P., Smith, G., and Stoker, G. (2009), 'Nudge Nudge, Think Think: Two Strategies for Changing Civic Behaviour', *The Political Quarterly* 80(3), pp. 361-370.

John, P., Cotterill, S., Moseley, A., Richardson, L., Smith, G., Stoker, G., and Wales, C. (2013), *Nudge, Nudge, Think, Think: Experimenting with Ways to Change Civic Behaviour (Revised Edition)*, London: Bloomsbury Academic.

John, P. (2014), 'Policy Entrepreneurship in UK Central Government: The Behavioural Insights Team and the Use of Randomized Control Trials', *Public Policy and Administration* 29(3), pp. 257-267.

Johnson, D.V. (2016), 'Twilight of the Nudges', *New Republic* (27th October 2016). Available at: https://newrepublic.com/article/138175/twilight-nudges.

Jolls, C. (2010), 'Dworkin's "Living Well" and the Well-Being Revolution', *Boston University Law Review* 90(2), pp. 641-656.

Jones, K. (2003), 'Emotion, Weakness of Will, and the Normative Conception of Agency', *Royal Institute of Philosophy Supplement* 52, pp. 181-200.

Jones, R., Pykett, J., and Whitehead, M. (2013), *Changing Behaviours: On the Rise of the Psychological State*, Cheltenham: Edward Elgar Publishing Limited.

Jung, J.Y., and Mellers, B.A. (2016), 'American Attitudes toward Nudges', *Judgment and Decision-Making* 11(1), pp. 62-74.

Kagan, S. (1989), *The Limits of Morality*, Oxford: Clarendon Press.

Kahneman, D. (2003), 'Maps of Bounded Rationality: Psychology for Behavioral Economics', *The American Economic Review* 93(5), pp. 1449-1475.

--------- (2011), *Thinking, Fast and Slow*, London: Penguin Books.

Kant, I. (2002), *Groundwork for the Metaphysics of Morals* (ed. and trans. by A.W. Wood), London: Yale University Press.

Keller, P.A., Harlam, B., Loewenstein, G., and Volpp, K.G. (2011), 'Enhanced Active Choice: A New Model to Motivate Behavior Change', *Journal of Consumer Psychology* 21, pp.376-383.

Kelly, J. (2013), 'Libertarian Paternalism, Utilitarianism, and Justice', in C. Coons and M. Weber (eds.), *Paternalism: Theory and Practice*, Cambridge: Cambridge University Press, pp. 216-230.

--------- (2014), 'The Life You Save May Not be Your Own', *The Good Society* 23(2), pp. 179-192.

Kelly, D., and Morar, N. (2016), 'Nudging and the Ecological and Social Roots of Human Agency', *American Journal of Bioethics* 16(11), pp. 15-17.

Korobkin, R. (2009), 'Libertarian Welfarism', *California Law Review* 97, pp. 1651-1685.

Korsgaard, C. (1983), 'Two Distinctions in Goodness', *The Philosophical Review* 92(2), pp. 169-195.

--------- (1986), 'Skepticism about Practical Reason', *The Journal of Philosophy* 83(1), pp. 5-25.

--------- (1997), 'The Normativity of Instrumental Reason', in G. Cullity and B. Gaut (eds.), *Ethics and Practical Reason*, Oxford: Clarendon Press, pp. 215-254.

Kosfeld, M., Heinrichs, M., Zak, P.J., Fischbacher, U., and Fehr, E. (2005), 'Oxytocin Increases Trust in Humans', *Nature* 435, pp. 673-676.

Kouchaki, M., and Desai, S.D. (2015), 'Anxious, Threatened, and Also Unethical: How Anxiety Makes Individuals Feel Threatened and Commit Unethical Acts', *Journal of Applied Psychology* 100(2), pp. 360-375.

Kramer, A.D.I., Guillory, J.E., and Hancock, J.T. (2014), 'Experimental Evidence of Mass-Scale Emotional Contagion Through Social Networks', *Proceedings of the National Academy of the Sciences of the United States of America* 111(24), pp. 8788-8790.

Kramer, M.H. (2017), *Liberalism with Excellence*, Oxford: Oxford University Press.

Krishnamurthy, M. (2015), 'Nudging Global Poverty Alleviation?', *The Law & Ethics of Human Rights* 9(2), pp. 249-264.

Kristjánsson, Kristján (2006), 'Habituated Reason: Aristotle and the 'Paradox of Moral Education'', *Theory and Research in Education* 4(1), pp. 101-122.

Lakoff, G., and Johnson, M. (2003), *Metaphors We Live By (New Edition)*, London: Chicago University Press.

Lavazza, A., and De Caro, M. (2010), 'Not so Fast: On Some Bold Neuroscientific Claims Concerning Human Agency', *Neuroethics* 3, pp. 23-41.

Leggett, W. (2014), 'The Politics of Behaviour Change: Nudge, Neoliberalism, and the State', *Policy and Politics* 42(1), pp. 3-19.

Levy, N. (2012), 'Ecological Engineering: Reshaping Our Environments to Achieve Our Goals', *Philosophy and Technology* 25(4), 589-604.

--------- (2017), 'Nudges in a Post-Truth World', *Journal of Medical Ethics* (online first: 19th May 2017). Available at: http://jme.bmj.com/content/early/2017/05/19/medethics-2017-104153

Lewis, H. (2016), 'Why the Author of "Nudge Theory" Is Worried His Theory Has Become Too Popular', *The New Statesman* (4th October 2016). Available at: www.newstatesman.com/politics/uk/2016/10/why-author-nudge-theory-worried-his-theory-has-become-too-popular.

Lillehammer, H. (2012), 'Autonomy, Value and the First Person', in L. Radoilska (ed.), *Autonomy and Mental Disorder*, Oxford: Oxford University Press, pp. 192-213.

MacMullen, I. (2013), 'Educating Children to Comply with Laws', *The Journal of Political Philosophy* 21(1), pp. 106-124.

Madva, A. (2016), 'Virtue, Social Knowledge, and Implicit Bias', in M. Brownstein and J. Saul (eds.), *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, Oxford: Oxford University Press, pp. 191-215.

MacKay, D., and Robinson, A. (2016), 'The Ethics of Organ Donor Registration Policies: Nudges and Respect for Autonomy', *The American Journal of Bioethics* 16(11), pp. 3-12.

Marteau, T.M. (1989), 'Framing of Information: Its Influence Upon Decisions of Doctors and Patients', *British Journal of Social Psychology* 28(1), pp. 89-94.

McDowell, J. (1998), *Mind, Value, and Reality*, Cambridge, MA.: Harvard University Press.

McKenna, M. (2008), 'The Relationship Between Autonomous and Morally Responsible Agency', in J.S. Taylor (ed.), *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, Cambridge: Cambridge University Press, pp. 205-234.

McKerlie, D. (2012), 'Alzheimer's Disease', in D. McKerlie, *Justice Between the Young and the Old*, Oxford: Oxford University Press, pp. 173-197.

McTernan, E. (2014), 'How to Make Citizens Behave: Social Psychology, Liberal Virtues, and Social Norms', *The Journal of Political Philosophy* 22(1), pp. 84-104.

Mele, A. (1995), *Autonomous Agents: From Self-Control to Autonomy*, Oxford: Oxford University Press.

Mele, A., and Shepherd, J. (2013), 'Situationism and Agency', *Journal of Practical Ethics* 1(1), pp. 62-83.

Mill, C. (1995), 'Politics and Manipulation', *Social Theory and Practice* 21(1), pp. 97-112.

Mill, J.S. (2005), *On Liberty*, New York: Cosimo Classics.

Miller, D. (1992), 'Deliberative Democracy and Social Choice', *Political Studies* 40, pp. 54 67.

Mills, C. (2013), 'Why Nudges Matter: A Reply to Goodwin', *Politics* 33(1), pp. 28-36.

--------- (2015), 'The Heteronomy of Choice Architecture', *Review of Philosophy and Psychology* 6, pp. 495-509.

Mitchell, G. (2005), 'Libertarian Paternalism Is an Oxymoron', *Northwestern University Law Review* 99(3), pp. 1245-1277.

Moles, A. (2015), 'Nudging for Liberals,' *Social Theory and Practice*, 41(4), pp. 644-667.

Moody-Adams, M. (2004), 'The Idea of Moral Progress', in C. Calhoun (ed.), *Setting the Moral Compass: Essays by Women Philosophers*, Oxford: Oxford University Press, pp. 256-272.

Moore, M. (2017), 'Society Will Be Defined By How We Deal With Tech Giants', *The Guardian* (2nd April). Available at: https://www.theguardian.com/commentisfree /2017/apr/01/brexit-britain-respond-tech-giants-civic-role-google-apple-facebook-amazon-eu.

Morewedge, C.K., and Kahneman, D. (2010), 'Associative Processes in Intuitive Judgment', *Trends in Cognitive Sciences* 14(10), pp. 435-440.

Mulhall, S., and Swift, A. (1996), *Liberals and Communitarians (Second Edition)*, Oxford: Blackwell Publishing.

Nagel, S.K., and Reiner, P.B. (2013), 'Autonomy Support to Foster Individuals' Flourishing', *The American Journal of Bioethics* 13(6), pp. 36-37.

Nagel, T. (1970), *The Problem of Altruism*, Oxford: Oxford University Press.

--------- (1986), *The View from Nowhere*, Oxford: Oxford University Press.

--------- (1991), *Equality and Partiality*, Oxford: Oxford University Press.

Niker, F. (2013), 'Transformative Nudging: Beyond Behaviour Change', MPhil thesis, University of Oxford.

--------- (forthcoming), 'Policy-led Virtue Cultivation: Can We Nudge Citizens towards Developing Virtues?', in T. Harrison and D. Walker (eds.), *The Theory and Practice of Virtue Education*, London: Routledge.

Niker, F., Reiner, P.B., and Felsen, G. (2015), 'Updating our Selves: Synthesizing Philosophical and Neurobiological Perspectives on Incorporating New Information into Our Worldview,' *Neuroethics* (published online: 18th December).

--------- (2016), 'Pre-authorization: A Novel Decision-Making Heuristic that May Promote Autonomy,' *The American Journal of Bioethics* 16 (5), pp. 27-29.

Noggle, R. (1996), 'Manipulative Actions: A Conceptual and Moral Analysis', *American Philosophical Quarterly* 33, pp. 43-55.

--------- (1999), 'Integrity, the Self, and Desire-Based Accounts of the Good,' *Philosophical Studies* 96(3), pp. 303-331.

Norman, D.A. (2013), *The Design of Everyday Things (Revised and Expanded Edition)*, Cambridge, MA.: MIT Press.

Nozick, R.(1974), *Anarchy, State and Utopia*, Oxford: Blackwell.

Nussbaum, M.C. (2001), *Upheavals of Thought: The Intelligence of Emotions*, Cambridge: Cambridge University Press.

Nye, S. (2012), 'Autonomy and Anti-Perfectionism', *Philosophy and Public Issues* 2(1), pp. 81-99.

O'Donnell, G. (2010), 'Applying Behavioural Insight to Health', London: Cabinet Office (31st December 2010).

OECD (2017), *Behavioural Insights and Public Policy: Lessons from Around the World*, Paris: OECD Publishing. Available at: http://www.oecd.org/gov/regulatory-policy/ behavioural insights-and-public-policy-9789264270480-en.htm.

Offer, A. (2006), *The Challenge of Affluence: Self-Control an Wellbeing in the United States and Britain Since 1950*, Oxford: Oxford University Press.

Olsaretti, S. (ed.) (2006), *Preferences and Well-Being*, Oxford: Oxford University Press.

Packard, V. (2007), *The Hidden Persuaders*, Brooklyn, NY.: Ig Publishing.

Parfit, D. (1984), *Reasons and Persons*, Oxford: Oxford University Press.

--------- (1997), 'Equality and Priority,' *Ratio* 10(3), pp. 202-221.

Parr, T. (2015), 'On the Job', PhD thesis (September 2015), University of Warwick. Available at: http://go.warwick.ac.uk/wrap/77848.

Pearson, A.R., Dovidio, J.F., and Gaertner, S.L. (2009), 'The Nature of Contemporary Prejudice: Insights from Aversive Racism', *Social and Personality Psychology Compass* 3, pp. 314-338.

Persson, I., and Savulescu, J. (2011), 'Unfit for the Future? Human Nature, Scientific Progress, and the Need for Moral Enhancement', in J. Savulescu, R. ter Meulen and G. Kahane (eds.), *Enhancing Human Capacities*, Oxford: Wiley-Blackwell, pp. 486-500.

--------- (2012), *Unfit for the Future: The Need for Moral Enhancement*, Oxford: Oxford University Press.

Plimmer, S. (2014), 'UK Cabinet Office 'Nudge' Team to be Spun Off into Private Group', *The Financial Times* (5th February). Available at: https://www.ft.com/content/571eef16-8d99-11e3-9dbb-00144feab7de.

Porter, T. (2011), 'Discussion Note: Colburn on Anti-Perfectionism and Autonomy', *Journal of Ethics and Social Philosophy*. Available at: http://www.jesp.org/articles/download/PorterDiscussion.pdf.

Posner, R.A. (1995), *Overcoming Law*, Cambridge, MA.: Harvard University Press.

Qizilbash, M. (2006), 'Well-Being, Adaptation and Human Limitations,' in S. Olsaretti (ed.), *Preferences and Well-Being*, Oxford: Oxford University Press, pp. 83-109.

--------- (2009), 'Well-Being, Preference Formation, and the Danger of Paternalism,' *Papers on Economics and Evolution* #0918.

--------- (2012), 'Informed Desire and the Ambitions of Libertarian Paternalism,' *Social Choice and Welfare* 38, pp. 647-658.

Quong, J. (2010), *Liberalism without Perfection*, Oxford: Oxford University Press.

Rachlinski, J.J. (2017), 'How I Learned to Stop Worrying and Love Nudges (Review of *The Ethics of Influence* by Cass Sunstein)', *Texas Law Review* 95(5), pp. 1061-1076.

Radoilska, L. (2012), 'Autonomy and Ulysses Arrangements', in L. Radoilska (ed.), *Autonomy and Mental Disorder*, Oxford: Oxford University Press, pp. 252-280.

Raihani, N.J. (2013), 'Nudge Politics: Efficacy and Ethics', *Frontiers in Psychology (Theoretical and Philosophy Psychology)* 4, pp. 1-3.

Railton, P. (1986a), 'Facts and Values,' *Philosophical Topics* 14(2), pp, 5-31.

--------- (1986b), 'Moral Realism,' *Philosophical Review* 95, pp. 163-207.

Raus, K., Focquaert, F., Schermer, M., Specker, J., and Sterckx, S. (2014), 'On Defining Moral Enhancement: A Clarificatory Taxonomy', *Neuroethics* 7, pp. 263-273.

Rawls, J. (1993), *Political Liberalism*, Cambridge, MA.: Harvard University Press.

--------- (1999), *A Theory of Justice (Revised Edition)*, Cambridge, MA.: Harvard University Press.

Raz, J. (1986), *The Morality of Freedom*, Oxford: Clarendon Press.

--------- (1989), 'Facing Up: A Reply', *Southern California Law Review* 62, pp. 1153-1235.

--------- (1994), *Ethics in the Public Domain: Essays in the Morality of Law and Politics*, Oxford: Clarendon Press.

--------- (1999), *Practical Reasons and Norms (Second Edition)*, Oxford: Oxford University Press.

--------- (2000), 'The Central Conflict: Morality and Self-Interest', in R. Crisp and B. Hooker (eds.), *Well-Being and Morality: Essays in Honour of James Griffin*, Oxford: Clarendon Press, pp. 209-238.

Rebonato, R. (2012), *Taking Liberties: A Critical Examination of Libertarian Paternalism*, London: Palgrave Macmillan.

Rees, C.F. (2016), 'A Virtue Ethics Response to Implicit Bias', in M. Brownstein and J. Saul (eds.), *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*, Oxford: Oxford University Press, pp. 191-214.

Riley, E. (2017), 'The Beneficent Nudge Program and Epistemic Injustice', *Ethical Theory and Moral Practice*, online first (published 5th April 2017, doi: 10.1007/s10677-017-9805-2).

Roberts, R.C. (2013), *Emotions in the Moral Life*, Cambridge: Cambridge University Press.

Rosati, C.S. (2016), 'Moral Motivation', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.

Rowson, J. (2011), 'Transforming Behaviour Change: Beyond Nudge and Neuromania', London: RSA.

Rutter, T. (2015), 'The Rise of Nudge – The Unit Helping Politicians to Fathom Human Behaviour', *The Guardian* (23rd July). Available online at: https://www.theguardian.com/public-leaders-network/2015/jul/23/rise-nudge-unit-politicians-human-behaviour.

Saghai, Y. (2013), 'Salvaging the Concept of Nudge,' *Journal of Medical Ethics* 39(8), pp. 487-493.

Samuels, R., Stich, S., and Bishop, M. (2012), 'Ending the Rationality Wars: How to Make Disputes about Human Rationality Disappear', in S. Stich, *Collected Papers, Volume 2: Knowledge, Rationality, and Morality, 1978-2010*, Oxford: Oxford University Press, pp. 191-223.

Sandel, M. (1994), 'Review of Political Liberalism', *Harvard Law Review* 107, pp. 1765-1794.

Sanders, M., and Halpern, D. (2014), 'Nudge Unit: Our Quiet Revolution is Putting Evidence at the Heart of Policy', *The Guardian* (3rd February). Available at: https://www.theguardian.com/public-leaders-network/small-business-blog/2014/feb/03/nudge-unit-quiet-revolution-evidence.

Savulescu, J., Douglas, T., and Persson, I. (2014), 'Autonomy and the Ethics of Biological Behaviour Modification', in A. Akabayashi (ed.), *The Future of Bioethics: International Dialogues*, Oxford: Oxford University Press, pp. 91-112.

Sawicki, N.N. (2016), 'Ethical Limitations on the State's Use of Arational Persuasion', *Law and Policy* 38(3), pp. 211-233.

Scanlon, T.M. (1975), 'Preference and Urgency,' *The Journal of Philosophy* 72(19), pp. 655-669.

--------- (1991), 'The Moral Basis of Interpersonal Comparisons,' in J. Elster and J.E. Roemer (eds.), *Interpersonal Comparisons of Well-Being*, Cambridge: Cambridge University Press, pp. 17-44.

--------- (1998), *What We Owe to Each Other*, Cambridge, MA.: Harvard University Press.

--------- (2010), *Moral Dimensions: Permissibility, Meaning, Blame*, Cambridge MA.: Harvard University Press.

Schubert, C. (2015), 'On the Ethics of Public Nudging: Autonomy and Agency,' Joint Discussion Paper Series in Economics, No. 33-2015.

Sen, A. (1987), *On Ethics and Economics*, Oxford: Basil Blackwell.

Service, O., Hallsworth, M., Halpern, D., Algate, F., Gallager, R., Nguyen, S.R., Sanders, M. (2014), 'EAST: Four Simple Ways to Apply Behavioural Insights' (11th April), London: BIT. Available at http://www.behaviouralinsights.co.uk/publications/east-four-simple-ways-to-apply-behavioural-insights.

Shafer-Landau, R. (2003), *Moral Realism: A Defence*, Oxford: Oxford University Press.

Shaw, T. (2017), 'Invisible Manipulators of Your Mind', *The New York Review of Books* (20th April). Available at: http://www.nybooks.com/articles/2017/04/20/kahneman-tversky-invisible-mind-manipulators.

Sher, G. (1997), *Beyond Neutrality: Perfectionism and Politics*, Cambridge: Cambridge: University Press.

Sherman, Nancy (1989), *The Fabric of Character: Aristotle's Theory of Virtue*, Oxford: Clarendon Press.

Shiffrin, S.V. (2000), 'Paternalism, Unconscionability Doctrine, and Accommodation', *Philosophy & Public Affairs* 29(3), pp. 205-250.

--------- (2004), 'Autonomy, Beneficence, and the Permanently Demented', in J. Burley (ed.), *Ronald Dworkin and His Critics*, Oxford: Blackwell Publishing, pp. 195-217.

--------- (2010), 'Inducing Moral Deliberation: On the Occasional Virtues of Fog', *Harvard Law Review* 123(2), pp. 1214-1246.

Shu, L.L., Mazar, N., Gino, F., Ariely, D., and Bazerman, M.H. (2012), 'Signing at the Beginning Makes Ethics Salient and Decreases Dishonest Self-reports in Comparison to Signing at the End', *Proceedings of the National Academy of Sciences* 109 (38), pp. 15197-15200.

Simon, H.A. (1990), 'Invariants of Human Behavior', *Annual Review of Psychology* 41, pp. 1–19.

Skinner, B.F. (1971), *Beyond Freedom and Dignity*, Harmondsworth: Penguin Books.

Slavny, A., and Parr, T. (2015), 'Harmless Discrimination', *Legal Theory* 21(2), pp. 100 114.

Smith, M.A. (1987), 'The Humean Theory of Motivation', *Mind* XCVI (381), pp. 36-61.

--------- (1994), *The Moral Problem*, Oxford: Blackwell.

Snow, N.E. (2010), *Virtue as Social Intelligence: An Empirically Grounded Theory*, London: Routledge.

Specker Sullivan, L., and Niker, F. (2017), 'Relational Autonomy, Maternalism, and the Nocebo Effect', *The American Journal of Bioethics* 17(5), pp. 52-54.

Standing, G. (2014), *The Precariat: The New Dangerous Class*, London: Bloomsbury.

Stanovich, K.E. (1999), *Who is Rational? Studies in Individual Differences in Reasoning*, Mahweh, NJ.: Erlbaum.

Stanovich, K.E. (2009), 'Distinguishing the Reflective, Algorithmic, and Autonomous Minds: Is It Time for a Tri-Process Theory?', in K. Frankish and J.B.T. Evans (eds.), *In Two Minds: Dual Processes and Beyond*, Oxford: Oxford University Press, pp. 55-88.

Stanovich, K.E., and West, R.F. (2000), 'Individual Difference in Reasoning: Implications For The Rationality Debate?', *Behavioural and Brain Sciences* 23, pp. 645–726.

Stears, M. (2005), 'The Vocation of Political Theory: Principles, Empirical Inquiry and the Politics of Opportunity', *European Journal of Political Theory* 4(4), pp. 325-350.

Stemplowska, Z., and Swift, A. (2012), 'Ideal and Nonideal Theory', in D. Estlund (ed.), *The Oxford Handbook of Political Philosophy*, Oxford: Oxford University Press, pp. 373-390.

Strauss, D.A. (1991), 'Persuasion, Autonomy, and Freedom of Expression', *Columbia Law Review* 91, pp. 334-371.

Sugden, R. (1993), 'Welfare, Resources, and Capabilities: A Review of *Inequality Reexamined* by Amartya Sen', *Journal of Economic Literature* 31, pp. 1947-1962.

--------- (2008), 'Why Incoherent Preferences Do Not Justify Paternalism,' *Constitutional Political Economy* 19(3), pp. 226-248.

Sunstein, C.R. (1995), 'Incompletely Theorized Agreements', *Harvard Law Review* 108(7), pp. 1733-1772.

--------- (1996), *Legal Reasoning and Political Conflict*, Oxford: Oxford University Press.

--------- (2013a), 'Deciding By Default', *University of Pennsylvania Law Review* 162(1), pp. 1-57.

--------- (2013b), 'The Storrs Lectures: Behavioral Economics and Paternalism', *The Yale Law Journal* 122(7), pp. 1826-1899.

--------- (2014a), *Why Nudge? The Politics of Libertarian Paternalism*, London: Yale University Press.

--------- (2014b), 'Nudges vs Shoves: The Benefits of Preserving Choice', *Harvard Law Review Forum* 127, pp. 210-217.

--------- (2015), *Choosing Not to Choose: Understanding the Value of Choice*, New York: Oxford University Press.

--------- (2016a), *The Ethics of Influence: Government in the Age of Behavioral Science*, Cambridge: Cambridge University Press.

--------- (2016b), 'People Prefer System 2 Nudges (Kind Of)', *Duke Law Journal* 66, pp. 121-168.

--------- (2017), 'Requiring Choice is a Form of Paternalism', *Journal of Behavioral Economics for Policy* 1(1), pp. 11-14.

Sunstein, C.R., and Reisch, L.A. (2014), 'Automatically Green: Behavioral Economics and Environmental Protection', *Harvard Environmental Law Review* 38, pp. 127-158.

Sunstein, C.R., and Thaler, R.H. (2003), 'Libertarian Paternalism is Not an Oxymoron', *The University of Chicago Law Review* 70(4), pp. 1159-1202.

--------- (2005), 'Libertarian Paternalism', in C.R. Sunstein (ed.), *Laws of Fear: Beyond the Precautionary Principle*, Cambridge: Cambridge University Press, pp. 175-203.

--------- (2006), 'Preferences, Paternalism, and Liberty,' in S. Olsaretti (ed.), *Preferences and Well-Being*, Oxford: Oxford University Press, pp. 233-264.

Swift, A. (2006), *Political Philosophy: A Beginners' Guide for Students and Politicians (Second Edition)*, Cambridge: Polity Press.

Tadros, V. (2011), *The Ends of Harm: The Moral Foundations of Criminal Law*, Oxford: Oxford University Press.

Tappolet, C. (2014), 'Emotions, Reasons, and Autonomy', in A. Veltmann and M. Piper (eds.), *Autonomy, Oppression, and Gender*, Oxford: Oxford University Press, pp. 163-180.

--------- (2016), *Emotions, Values, and Agency*, Oxford: Oxford University Press.

Taylor, J.S. (2008), 'Introduction', in J.S. Taylor (ed.), *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, Cambridge: Cambridge University Press, pp. 1-29.

Taylor, A. (2017), 'Denmark Is Naming an Ambassador Who Will Just Deal With Increasingly Powerful Tech Companies', *The Washington Post* (4th February). Available at: https://www.washingtonpost.com/news/worldviews/vwp/2017/02/04/denmark-is-naming-an-ambassador-who-will-just-deal-with-increasingly-powerful-tech-companies.

Terbeck, S., Kahane, G., McTavish S., Savulescu, J., Cowen, P., and Hewstone, M. (2012), 'Propranolol Reduces Implicit Negative Racial Bias', *Psychopharmacology* 222, pp. 419–424.

Terranova, T. (2012), 'Attention, Economy and the Brain', *Culture Machine* 13, pp. 1-19.

Thaler, R.H., (2015), 'Foreword', in D. Halpern, *Inside the Nudge Unit: How Small Changes Can Make a Big Difference*, London: W.H. Allen, pp. ix-xiv.

Thaler, R.H., and Sunstein, C.R. (2009), *Nudge: Improving Decisions about Health, Wealth and Happiness*, London: Penguin Books.

Theotokis, A., and Manganari, E. (2015), 'The Impact of Choice Architecture on Sustainable Consumer Behavior: The Role of Guilt', *Journal of Business Ethics* 131, pp. 423-437.

Trout, J.D. (2005), 'Paternalism and Cognitive Bias', *Law and Philosophy* 24(4), pp. 393-434.

Tsai, G. (2014), 'Rational Persuasion as Paternalism', *Philosophy & Public Affairs* 42(1), pp. 78-112.

Tversky, A., and Kahneman, D. (1974), 'Judgment under Uncertainty: Heuristics and Biases', *Science* 185(4157), pp. 1124-1137.

United Nations (2016), 'Behavioural Insights at the United Nations' (20th December). Available at: http://www.undp.org/content/undp/en/home/librarypage/development-impact/behavioural-insights-at-the-united-nations--achieving-agenda-203.html.

Upton, C.L. (2009), 'Virtue Ethics and Moral Psychology: The Situationism Debate', *The Journal of Ethics* 13(2), pp. 103-115.

Van Ijzendoorn, M., and Bakermans-Kranenburg, M. (2011), 'On Embodied and Situational Morality', in D. de Ruyter and S. Miedema (eds.), *Moral Education and Development*, Rotterdam: Sense Publishers, pp. 13-30.

Verbeek, P-P. (2009), 'Ambient Intelligence and Persuasive Technology: The Blurring Boundaries Between Human and Technology', *Nanoethics* 3, pp. 231-242.

Waldron, J. (1994), 'Disagreements about Justice', *Pacific Philosophical Quarterly* 75, pp. 372-387.

--------- (2015), 'It's All For Your Own Good', *The New York Book Review* (9th October). Available at: http://www.nybooks.com/articles/2014/10/09/cass-sunstein-its-all-your-own-good.

Wall, S. (1998), *Liberalism, Perfectionism, and Restraint*, Cambridge: Cambridge University Press.

--------- (2012), 'Perfectionism in Moral and Political Philosophy,' in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.

--------- (2013), 'Moral Environmentalism', in C. Coons and M. Weber (eds.), *Paternalism: Theory and Practice*, Cambridge: Cambridge University Press, pp. 93-114.

Wallace, R.J. (1999), 'Three Conceptions of Rational Agency', *Ethical Theory and Moral Practice* 2(3), pp. 217-242.

Wansink, B., and van Ittersum, K. (2013), 'Portion Size Me: Plate-Size Induced Consumption Norms and Win-Win Solutions for Reducing Food Intake and Waste', *Journal of Experimental Psychology (Applied)* 19(4), pp. 320-332.

Watson, G. (1975), 'Free Agency', *The Journal of Philosophy* 72(8), pp. 205–220.

Watzlawick, P., Weakland, J., and Fisch, R. (1974), Change: Principles of Problem Formation and Resolution, New York: W.W. Norton.

Wertheimer, A. (2015), 'Coercion', Wiley Encyclopedia of Management 2, pp. 1–2.

White, M.D. (2013), *The Manipulation of Choice: Ethics and Libertarian Paternalism*, New York: Palgrave Macmillan.

Whitehead, M., Jones, R., and Pykett, J. (2011), 'Governing Irrationality, or More than Rational Government? Reflections on the Rescientisation of Decision Making in British Public Policy', *Environment and Planning A* 43(12), pp. 2819-2837.

Wieland, J.W. (2016), 'Responsibility for Strategic Ignorance', *Synthese* (published online first: 29th June; doi: 10.1007/s11229-016-1145-6).

--------- (2017), 'Willful Ignorance', *Ethical Theory and Moral Practice* 20(1), pp. 105-119.

Wiggins, D. (1987), 'A Sensible Subjectivism?', in D. Wiggins, *Needs, Values, Truth: Essays in the Philosophy of Value*, Oxford: Blackwell, pp. 185-211.

Wilkinson, T.M. (2013), 'Nudging and Manipulation,' *Political Studies* 61(2), pp. 341-355.

--------- (2016), 'Counter-Manipulation and Health Promotion', *Public Health Ethics* (online first: 14th November; doi: 10.1093/phe/phw044).

Williams, J. (2013), 'Staying Free in a World of Persuasive Technologies', *Oxford Internet Institute Blog* (29th July). Available at: https://www.oii.ox.ac.uk/blog/ staying-free-in-a-world-of-persuasive-technologies.

Wolf, S. (2008), 'Freedom within Reason', in J.S. Taylor (ed.), *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, Cambridge: Cambridge University Press, pp. 258-274.

Wolff, J. (2011), *Ethics and Public Policy: A Philosophical Inquiry*, Abingdon: Routledge.

Woodward, J. (1986), 'The Non-Identity Problem', *Ethics* 96(4), pp. 804-831.

World Bank (2015), *Mind, Society, and Behavior* (World Development Report 2015), Washington, DC.: World Bank. Available at: http://www.worldbank.org/en/ publication/wdr2015.

Wu, T. (unpublished), 'Attention Markets and the Law (26th March 2017)'. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2941094.

Zagzebski, L. (2003), 'Emotion and Moral Judgement', *Philosophy and Phenomenological Research* 66(1), pp. 104-124.

Zarpentine, C. (2013), "The Thorny and Arduous Path of Moral Progress': Moral Psychology and Moral Enhancement', *Neuroethics* 6, pp. 141-153.