Institut d'études politiques de Paris

ECOLE DOCTORALE DE SCIENCES PO

Master's in Political Science – Programme in Political Theory

Democracy for the Cognitively Vulnerable

Nudging in Liberal Democracies

Barış Can Kaştaş

Under the supervision of Annabelle Lever (Professeure in Sciences Po)

Abstract

This dissertation will tackle the topic of cognitive biases and their use as a political tool within the context of democratic theory. It takes cognitive biases, as they are defined and codified by cognitive psychology, to be an inevitable part of every human being's mind, and thus a factor be considered in the political sphere. These cognitive biases follow a tradition in political theory that has treated questions of rational defects before, but whose approach and solutions were quite different from contemporary understanding of cognitive biases (the dissertation will focus on the cognitive defects observed by empiricists and utilitarians in early modern philosophy as a point of comparison). These biases have come into the attention of a larger public in the late 2000s, accompanying a "behaviourist shift" in economics and public policy that deviated from classical theories. This change of models in economics and public policy brought the potential use of cognitive policies as political tools to the forefront, with the notable influence of *Nudge Theory* in the USA and the UK as chief examples. The inclusion of cognitive biases to the lexicon of liberal democracy poses a *dilemma*: on the one hand, the unfettered use of cognitive biases in a political context proves to be more problematic for liberal democracies than what libertarian paternalists propose On the other hand, the revelation that political agents are predictably irrational has consequences that theory cannot ignore. This double dilemma calls for a new conception of democracy: a democracy with cognitively *vulnerable* subjects means the core tenets of democracy must be weighed against our fallible nature, with a fallible (but improvable) subject at its core.

Acknowledgements & Dedication

This dissertation is the result of guidance from several different sources that I hope to have put to good use. I am forever grateful for their contribution, and would like to dedicate some space to thanking them.

I'd like to thank above all Annabelle Lever, my supervisor, for her invaluable help throughout the writing process. I'm also grateful to Gloria Origgi, who guided me during the initial months of my preparation, but with whom I sadly couldn't work until the end.

I am also thankful for the input offered by my fellow Master's students during our writing workshop in this semester, as well as Astrid von Busekist and Frédéric Gros. Their feedback following my presentation in the workshop, as well as suggestions from classmates for new sources, have added an invaluable dimension to my project.

This dissertation is dedicated to my parents, Ümit & İrfan Kaştaş, whose debt I can never repay, and *Volpes Umbrae*, my companions in adventure during the Master's – and beyond.

Table of Contents

Introduction

This dissertation will tackle the topic of cognitive biases and their use as a political tool within the context of democratic theory. It takes cognitive biases, as they are defined and codified by cognitive psychology, to be an inevitable part of every human being's mind, and thus a factor be considered in the political sphere. These cognitive biases follow a tradition in political theory that has treated questions of rational defects before, but whose approach and solutions were quite different from contemporary understanding of cognitive biases (the dissertation will focus on the cognitive defects observed by empiricists and utilitarians in early modern philosophy as a point of comparison). These biases have come into the attention of a larger public in the late 2000s, accompanying a "behaviourist shift" in economics and public policy that deviated from classical theories. This change of models in economics and public policy brought the potential use of cognitive policies as political tools to the forefront, with the notable influence of *Nudge Theory* in the USA and the UK as chief examples. The dissertation will examine the consequences of this historical development to normative democratic theory, and liberal theories of democracy in particular.

The main argument presented by the dissertation is the following: The inclusion of cognitive biases to the lexicon of liberal democracy poses a *dilemma*: on the one hand, the unfettered use of cognitive biases in a political context proves to be more problematic for liberal democracies than what libertarian paternalists propose On the other hand, the revelation that political agents are predictably irrational has consequences that theory cannot ignore. Unlike what the proponents of *libertarian paternalism* may claim, the underlying assumptions about human rationality in cognitive biases challenge key assumptions of liberal democratic conceptions of politics. Theories of liberal democracy will thus need to reconcile the scientific findings with their key assumptions about individual agents. The focus on the use of cognitive biases by *choice architects* who can help individuals achieve *their own goals* calls into question what individual welfare and truth mean in a democracy, in a manner that is diametrically opposed to the ignorance of cognitive biases: while incorporating biases means classic conceptions of individual preference become meaningless (thus undoing the legitimacy of liberal polities), ignoring them invites the very practical problem of third-party abuse.

This double dilemma calls for a new conception of democracy: a democracy with cognitively *vulnerable* subjects means the core tenets of democracy must be weighed against our fallible nature, with a fallible (but improvable) subject at its core. This potential for improvement offers a natural critique of libertarian paternalism and other proposals for the use of cognitive biases for policy. A *democratisation* of these proposals, where specific democratically-decided ends for which the exploitation of citizens' cognitive biases by the government would be justified, offers to be an alternative to libertarian paternalism.

A - Definitions and Theoretical Assumptions

In order to avoid losing itself in a specialised jargon that nobody except the author of this dissertation uses, terms within this dissertation will be used in their most common understanding unless otherwise specified. With that said, underlining some notable properties of several terms will be useful when these properties come into play during later chapters. While definitions are offered here without any context, the logical development of the ideas within this dissertation will hopefully present any necessary context as needed.

First of all, all political concepts tackled by this dissertation are understood within the context of **liberal democracies**. Liberal democracies can be understood as the combination of two specific strands of political thought: *liberalism* and *democracy*. Liberalism, in this context, refers to the belief that every person is *free* and *equal*. This demand for freedom and equality is usually interpreted as the need for everyone to be given *equal moral weight*. In other words, everyone is *treated* equally. The exact way in which equal treatment is given differs from thinker to thinker (e.g. the fair distribution of primary goods in Rawlsian liberalism, an ambition-sensitive and endowment-insensitive system of auctions and insurances in Ronald Dworkin, John Roemer's various proposals to offset inequalities between different social strata etc.), but this key tenet seems to be accepted by every liberal thinker[1]. A more concrete and common interpretation of this freedom and equality takes the form of equality before law, and a set of inviolable fundamental rights for every person that takes precedence over other rights. Democracy, on the other hand, has taken many different forms throughout history, but is usually summed up as "rule by people". The common point to every government that claimed the term

[1] Will Kymlicka. *Contemporary Political Philosophy*, p. 53-101.

for themselves is the participation of the entire *demos* in some form or another (through elections, plebiscites, deliberations, etc.) Thus, democracy for our purposes is defined as the *participation* of every citizen in political decisions that concern them, with the exact scope of participation and decisions remaining undefined.

Liberal democracy, then, is the combination of these two ideas: it denotes all political systems that meet the standards set by liberal *and* democratic principles. In historical terms, this has referred to Western governments with free and regular elections for public office, a constitution that guarantees fundamental rights of individuals, and a rule of law to delimit the use of public authority. Such a political system is thus considered as the default assumption[1] on which every other concept is introduced in this dissertation. Individuals in a liberal democracy (or any political system for that matter) can be deemed as **rational** or **irrational** depending on whether their preferences align with the rational decision-making processes within them. While different structures can be imagined for such decision-making processes, they tend to measure an action or decision's rationality on whether the action is coherent with the situation in which the agent believes to take that action, and with the aimed result of the action itself.

An action can be deemed as irrational primarily due to two factors: if the decision-making process is affected by **passions**, or if a **cognitive defect** causes the reasoning to be faulty. Cognitive defects are separate from passions in that passions deal with affective detriments to rationality while cognitive defects, for this dissertation, are concerned with deviations from rationality with an observable pattern. This means that both passions and cognitive defects bring about the same result (irrationality), but due to different reasons: passions overtake decision-making in light of a strong emotion or desire, often in an unpredictable manner, whereas cognitive defects are mental shortcuts that reliably lead to a certain faulty result.

Nudges, which are a novel form of dealing with cognitive defects with which the majority of the dissertation will be concerned, can be understood as remedies to cognitive defects. They can be categorised into three groups depending on how they remedy said defects:

[1] As we will see in later chapters, default assumptions can also hide dangerous cognitive biases in themselves. But in order to have a *primum movens*, we will leave the criticism of this default assumption outside this dissertation.

trainings that aim to remove a cognitive defect entirely from a decision making process, *incentives* that aim to make rational outcomes more preferable to overcome the overvaluing of certain choices due to cognitive defects, and Nudges that *exploit one defect against another*, so that the agent still follows a faulty line of reasoning to nevertheless arrive at a rational conclusion. **Libertarian paternalism** is a specific ideology, devised by the proponents of contemporary Nudge theory (primarily Richard Thaler and Cass Sunstein) that justified the use of Nudges on two principles: firstly, that Nudges should not actually limit choices in decision-making but only influence agents towards certain options. Secondly, that agents should be Nudged towards choices that will increase their welfare, where their welfare is judged by themselves[1].

Having presented the definitions on which this dissertation's topic stands, I feel the need to approach this groundwork once again from an outside perspective. While I believe the definitions I have provided are sufficiently well-developed to be used to treat the topic as a whole, they are developed only as a bridge between the specific fields this dissertation tackles. There remains a conceptual universe that is outside these fields, and that could call into question the common assumptions that these definitions had simply taken for granted. We should give due consideration to these assumptions, either justifying them or defending them against potential criticisms.

The core assumptions that come from the fields I choose are of course the paradigms that created them. In the case of psychological bases, a dissertation that treats cognitive biases has to follow a cognitivist paradigm, which stands in stark contrast to behaviourist or other psychological paradigms. This means that I admit that the inner processes of the mind can be studied, and that they act as a bridge to external environment and stimuli, with their own effects on human behaviour. The concept of cognitive biases, let alone their political use, would be simply absurd if no inner structure that can limit an individual's agency can be admitted. This is thus one core assumption that the dissertation holds to be true, without justifying it extensively.

Similarly, the primacy of the individual as an agent is a core assumption that stems from the psychological and economic sources of this project. While I tread carefully in order to avoid

[1] Richard Thaler & Cass Sunstein, *Nudge*, p. 4-5.

describing the individual as a monad that is completely independent of its environment, I still use the individual as the basic building block of this dissertation, on which every other piece stands. I believe this assumption makes intuitive sense: we all experience life as individuals, and while these individuals can be defined outside their inner and outer structures with difficulty, they are still *individuals* within those structures.

But the way in which this individual is defined also brings its own assumptions. The economic and psychological sources of this dissertation mean that actions of individuals are assumed to be identified in their patterns at least to a certain extent. The exact extent to which this is possible, however, raises some questions. These questions become evident in the definition of rationality: this dissertation defines rationality as selecting the most efficient means for a given end of an agent's preference. Thus, I maintain a framework of efficiency calculation. While early modern philosophers and cognitive scientists treated in the first two chapters would be perfectly fine with this definition, some criticism is still possible. There's notably some objections to the primacy of cognitive models of explaining human behaviour. These objections claim that what cognitive models define as "irrational" and as "error" are simply conclusions drawn from different epistemological ground rules, that the elevation of rationality as the source of correct action and emotion as a disruptor of our "true" will that leads us astray is a loaded definition that commits an epistemic injustice to other ways of decision-making. For the moment, I have no satisfactory answer against these objections. I believe a study of what it truly means to be irrational, in which the context of emotions and cognitive defects can be redefined, would be a very interesting endeavour. However, it goes far beyond the scope of this dissertation. As a result, I will be taking for granted the idea that rationality is finding the most efficient way to achieve a given end, and all the negative outlook on emotions and cognitive defects this brings.

B - Historical Context

The terms just described came to be appreciated as such within the existing literature only after a long development. This development in itself was made possible by the historical intersection of different fields, mainly behavioural psychology, economics and finally public policy.

The idea that humans are not fully rational (thus, at the mercy of specific cognitive defects) is almost as old as the notion of rationality itself. While classical philosophy tied this irrationality to the competition between reason and emotions as the causes of human action, causes for irrationality beyond passion started to be considered with early modern philosophy: empiricists and utilitarians alike noted the tendency of humans to prefer avoiding pain to gaining pleasure, that they had a tendency to think in the short term to the detriment of the long term, or that their reason erred in other similar ways requiring correction. This was a departure from the reason/passion dichotomy as it named a reason for rationality to fail that was beyond pure emotion, in ways that are very similar to our modern understanding of **biases**. Nevertheless, the conception of an objective and universal reason, which naturally aligned with humanity's rational faculties, meant that human rationality's defects were not fully explored.

The question of cognitive biases on its own surfaced following a cognitivist turn in the 60s in psychology. The dominant line of thinking in the field previously was **behaviourism**[1], which attempted to understand the workings of the human mind as a set of behaviours, which were themselves set as responses to environment or past stimuli. In behaviourism, the environment and the past of an individual directly shaped the mind of the individual, causing its actions to take a certain direction over another.

Cognitive psychology, however, marked a shift in the way the mind was conceptualised[2]. Unlike behaviourism, which attempted to study the effects of external stimuli and behaviour and took little interest in the inner reflection of these concepts, cognitive psychology approached the mind as a tool of *cognition* – in other words, of *knowing*. Cognitivist approaches thus started examining the mental processes through which external stimuli and behaviour were perceived. In a cognitivist vision of the mind, the mind was no longer directly shaped by external stimuli and environment – these external factors passed through the filter of mind's inner processes, which had their own qualities worth studying.

One such quality researchers focused on was the mind's deviation from rational choices in a recognisable pattern. The psychologists Daniel Kahneman and Amos Tversky introduced

[1] George Mandler, *Origins of the cognitive (r)evolution*.

[2] *Ibid.*

the notion of cognitive biases in 1972[1], through their comparison of cognitive models of the mind with economic models of choice. This intersection between psychology and economics challenged the classical paradigm of human decision-making economics had kept since its classical period. Previous theories had assumed individual agents to be rational. Economists as early as the 19th century defined humans as an *economic man*, who made choices according to his interests from different outcomes, which he calculated rationally. The term reached prominence in the beginning of the 20th century through the rational choice theory of British economist Lionel Robbins[2], and the paradigm of *homo economicus* shaped the field of economics as one where individuals calculated gains and losses, and acted according these alone. The new cognitivist critique instead painted a "predictably irrational" model of human behaviour, with cognitive biases explaining how these predictable irrational patterns functioned. This criticism of the classic economic models led a deeper study of psychological and emotional components in individual decision-making, spawning what is now called *behavioural economics*[3].

While debates within psychology and economics continued on the validity of cognitive models as well as their exact content continued, political applications of these models became a topic of interest by the end of the 90s. The term *nudge*, coined by James Welk before 1995[4], was catapulted into prominence in 2008 by Richard Thaler and Cass Sunstein's book *Nudge: Improving Decisions About Health, Wealth and Happiness*, which placed the term in the specific ideological context of *libertarian paternalism*. Thaler and Sunstein transformed the theoretical insights of behavioural economics (and its inspirations from cognitive psychology) into concrete policy proposals, as well as a specific design philosophy that guided these proposals.

The specific context in which nudges (and the use of cognitive biases instead of usual policy tools) gained traction as a political strategy should be noted: a world economy still in the throes of the Great Recession, with the majority of governments in Europe and North America

[1] Daniel Kahneman & Amos Tversky, *Subjective probability: A judgement of representativeness*.

[2] Lionel Robbins, *An Essay*, p. 87.

[3] Which, ironically was inspired by a school of psychological thought that opposed *behaviourism* in psychology.

[4] James Welk, *Mind, nature and the emerging science of change*.

electing to pursue austerity measures to curb the crisis, meant that the political scene was on the lookout for a cost-effective policy tool that promised to be just as effective as classic subsidies and welfare programmes. The ideology of libertarian paternalism proved to be a great fit for this demand, with governments in the UK and the USA giving a primary spot to nudges in their executive design. The British government under Prime Minister David Cameron created Behavioural Insights Team (colloquially known as the *Nudge Unit*) within the Cabinet Office, while American President Barack Obama placed Cass Sunstein himself as an Administrator to the Office of Information and Regulatory Affairs.

The interest generated in the 2000s by the use of cognitive biases for welfare improvement did not last long, however. The enthusiasm for nudges and their beneficial effects was suddenly replaced with the changing political atmosphere: the defeat of US Democrat presidential nominee Hillary Clinton (whose platform was, on many ways, a continuation of Obama's policies) to Donald Trump, as well as the resignation of David Cameron after the successful Brexit vote (which he campaigned against) in 2016 served as tell-tale signs of ideological change. This change of political atmosphere took a sinister tone when its conditions[1] were revealed: the Cambridge Analytica scandal showed the extent to which private interests used methods such as narrowcasting[2] and other appeals to the public's *cognitive defects* to ensure the success of Donald Trump and the Leave outcome respectively. The methods used by Cambridge Analytica marked a paradigm shift: the cognitive biases that Nudge theory had pinpointed could be used for far more nefarious ends. In a way, this revelation changed the public's approach to cognitive biases: the focus turned less on the use of these biases to improve individual welfare, and more to their discussion as (as previously stated) cognitive *vulnerabilities*, and on ways to circumvent them through information and stronger incentives (what is sometimes called *debiasing*).

[1] These conditions undoubtedly include the normalisation of xenophobic and authoritarian ideas (the rise of the so-called "alt-right"), as well as the erosion of trust in traditional political institutions and a common conception of truth, but for now we will be primarily interested in the misuse of cognitive biases during this ideological shift.

[2] In this context, narrowcasting refers to the practice whereby online advertisers aim their content to *specific* groups. These groups are defined and created by data collected about users that put them into narrow categories, allowing them to be efficiently targeted.

I - Cognitive Defects – New and Old

While the discussion of cognitive deficiencies in political subjects has existed for a considerable amount of time in political theory, the way in which the question was treated has changed in the contemporary era. This chapter provides an account of the "classical" way cognitive defects and biases have been treated, and compares it with the contemporary discussions of cognitive biases. Such a comparison shows how classical accounts of cognitive defects focus on two errors: errors of ends – people choosing the wrong goals, where the answer is usually education; or errors of means – people using the wrong tools regardless of the ends they choose, which require a workaround as a solution.

A - Classical Cognitive Defects

A particular brand of modern philosophy paid special attention to the cognitive shortcomings of human beings: British empiricists and later utilitarians[1] created accounts of human reason that was based on observations of human behaviour, and utilitarians notably aimed to calculate the moral worth of certain actions over others by a calculus of utility. This calculus approach inevitably produced some criticism of human tendencies when they produce a shift in behaviour compared to what is optimal, in much the same way as modern studies on cognitive biases did.

This section will attempt to demonstrate that early modern philosophers (notably John Locke, John Stuart Mill and Jeremy Bentham) approached the question of cognitive defects as a problem of ends or means. As such, the proposed solutions are also corrections of ends or means in themselves – either by making self-interest align with public interest through the design of political institutions, or by developing methods of ethology that allow people to improve their character, or by taking their defects of means into account when deciding on a political course of action.

[1] While I by no means mean that there is a straight line between these two different schools of thought, I believe a certain continuity can be found. Most notably, the empiricist desire to build all philosophy from observable facts is quite compatible with the utilitarian desire to base ethics on practical grounds of utility – utility which is usually observed through sensory data.

1) Defects of Means versus Defects of Ends

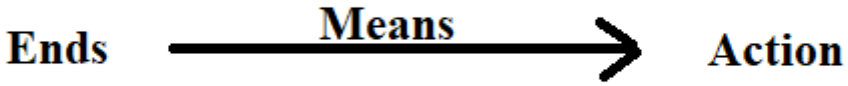
An important distinction can be made when discussing types of rationality and irrationality: whether we are judging an action as rational or irrational due to its ends – i.e. whether the actor had set a reasonable goal to achieve – or due to its means – i.e. whether the actor's actions aligned with the end they had set themselves, and helped achieve it[1]. The basis on which the rationality of an action can be judged differs considerably depending on which of these frames we're applying.

The rationality of ends concerns itself with the validity of an end regardless of the means that would be used to achieve it. It is simply asking the question of whether the destination is worth any journey at all – regardless of what said journey may entail. As such, its considerations are less focused on efficiency and deal instead with the legitimacy. This legitimacy is often discovered through moral intuitions. These moral intuitions often take the form of some first principles, axioms that seem evident, and from which a whole body of moral deductions can be made. Cognitive defects, then, arise when people fail to arrive at some moral intuition when they *should* have been capable of doing so. Attempting to understand whether the demands of a group of protestors are reasonable would be an exercise of ends, as it tries to understand what end state they wish to bring about – whether what they defend is just or unjust.

In comparison, the rationality of means asks whether the course of action for any given end (without discussion on the validity of the end itself) is truly appropriate. This appropriateness can be asked in a strong manner – whether the action chosen is truly in line with the goal, given its results, assumptions, and content – or in a weaker manner – whether the action chosen is the most *efficient* way of achieving said goal, in other words, whether the action maximises the result that was aimed. Asking whether the various methods the same protestors employ – marching, civic disobedience, rioting etc. – help them get closer to their goal is then asking the rationality of their means.

Thus, a means/ends based theory of action seems to follow a simple diagram:

[1] This ends/means distinction is in fact a younger distinction compared to the classical ideas we will be discussing, and can be traced to Max Weber's *zweckrational/wertrational* distinction, or Rawls' distinction between "the rational" and "the reasonable". Thus, its earliest form can be traced to early 20th century. Nevertheless, we can make the argument that ideas of cognitive defects proposed by empiricists and utilitarians fit in the framework that came after them.



An individual sets an end, and decides on some means to achieve it. These two decide the specific course of action the individual will take, and if neither the end, nor the means are irrational, then the action taken is considered rational.

Of course, the examination of an action's rationality doesn't have to focus on *only* means or *only* ends. In fact, the two can sometimes be linked. The ends chosen might sometimes oblige them to pursue a specific path, or make certain methods unacceptable (a pacifist revolution rioting with the intention to kill is a contradiction). Similarly, certain means might make certain ends inevitable (a violent suppression of protests often leads to the escalation of violence and people getting radicalised, for example). Thus, while these categories are useful for judging the rationality of an action, it becomes clear that they are not mutually exclusive spheres of analysis. As we will see, however, the solution of defects in one sphere was limited to that sphere in early modern philosophy.

2) Ends-Based Defects, or Faults of Character

Having made the distinction between different kinds of rational analysis, we can now see what early modern philosophers have said about cognitive defects, starting with ends-based defects.

The first philosopher whose ideas we shall take into consideration is John Locke. Locke's thoughts on the limits of human understanding, which can be found conveniently in *An Essay Concerning Human Understanding*, seem to be thus: while humans are capable of understanding the essence of certain concepts (such as mathematical ideas or God), we only

have a sensory perception of things pertaining to the real world[1]. Thus, Locke claims that our understanding of anything of practical importance is based on probability[2].

This incomplete nature of human understanding lead Locke to place a great emphasis on education as a corrective tool. He argued that parents had the responsibility to educate their children in a way that developed rationality. In *Some Thoughts Concerning Education* and *The Conduct of Understanding*, Locke clarifies that the development of reason is necessary for the individual to make fair judgements, and take correct actions. Without such a development, reason in Locke's view is clouded by partiality, intolerance and passion. The cognitive defects that beset an underdeveloped reason are thus the ones we just mentioned.

What is the common characteristic of these three defects? The context in which they were defined is important to note: Locke sees partiality, intolerance and passion as an impediment to making fair judgements, and taking correct actions[3]. In other words, for Locke, the cognitive defects of an individual exist in so far as they can cause a false judgement in *value*, in the justness of their action. We thus see that Locke's account of cognitive defects is concerned primarily with the rationality of *ends*. His solution to the individual erring in the destination they should pick is an education that helps them develop their reason, and thus avoid the pitfalls he has identified. This education proves to be a solution for Locke because it promotes the "correct" values (or rather, the correct ends) by inculcating "justice as respect for the rights of others, civility, liberality, humanity, self-denial, industry, thrift, courage, truthfulness, and a willingness to question prejudice, authority and the biases of one's own self-interest"[4].

Going from empiricism to utilitarianism, the works of Jeremy Bentham give some consideration to cognitive defects as well. Much like Locke, Bentham's understanding of a cognitive defect is concerned with the ends individuals choose for themselves. In effect,

[1] John Locke, *An Essay Concerning Human Understanding*, Book IV.14.1–2, N: 652

[2] Here, the term shouldn't be understood as a mathematical calculus, as Locke – as far as historians can tell – was unaware of the budding science of probability at his time. Probability instead is used in the sense of a *probable* judgement, one whose Proofs are not constant or immutable, but nevertheless convince the mind of their correctness, only to a lesser extent.

[3] *idem*, 1996: xii

[4] *ibid*: xiii

Bentham posits as a given that there may be a disjunction between an individual's "real" interests and their "perceived" interests. This disjunction poses a particular problem for Bentham's political theory, where political action is based on public interest, which is simply the aggregation of individual interests[1]. As a result, the legislator is given the very important task of knowing what the "real" interests of the governed are, and lead them from their perceived interests to these correct goals. The way to achieve this complex goal is indirect law, which Bentham defines as "a secret plan of connected and long-concerted operations to be executed in the way of a stratagem or *petite guerre*"[2]. These complex operations would, besides barring certain behaviours for individuals, also motivate them to act in several more desirable ways, through "codes of behaviour", moral instruction, teaching of history, biographies and even works of art could play a role in this "soft power"[3].

Another political concern of Bentham that is less strongly tied to cognitive defects was the question of self-interest[4]. Bentham believed that political institutions needed to be designed in a way that allowed the personal interests of those holding office to align with the public interest. More concretely, he argued that political institutions should *not* be based on a fixed formula of institutional balance, but instead allow easy changes between the governed and the governors so that their interests could blend[5]. A notable conclusion to draw from this consideration of self-interest is the following: while Bentham follows the ends-based critique of cognitive defects of Locke, his suggestion to *design around* these defects (either in the case of self-interest or in the case of indirect law to promote public interest) seems to be departure in terms of solutions (one that is notably closer in spirit to nudging), while still solving ends-based defects through the modification of ends.

The last philosopher whose ideas on cognitive defects are worthy of examination is John Stuart Mill. Mill's ideas are formed primarily as a response to Bentham's utilitarianism, mostly

[1] Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation*, p. 12

[2] *idem*, *Of the Limits of the Penal Branch of Jurisprudence*, p. 233

[3] *idem*, *Bentham Manuscripts at University College London*, lxxxvii, p. 18–19

[4] It could be argued that self-interest in public servants, due to failing to maximise public utility for personal benefit, would be a cognitive defect under utilitarian principles, since on an ends-based analysis, that is the reasonable (in Rawlsian terms) course of action.

[5] *idem*, *A Comment on the Commentaries and A Fragment on Government*, p. 485

by differentiating his own interpretation of utilitarianism in the book simply titled *Utilitarianism*. Mill is primarily critical of Bentham's view of humans as entirely selfish: Mill believes Bentham's model fails to account for cases of altruism, nor the effect the environment has on human character. Namely, the idea of utilitarianism as a useful framework becomes questionable if everyone acts in their self-interest. After all, nobody would be implementing it if they only looked after their self-interest. Mill's answer to this dilemma comes from his vision of human nature as extremely pliable: A proof of their pliability is their variation between different cultures, as Mill firmly believes that human nature is "astonishingly pliable"[1]. This pliability leads to a specific vision of cognitive defects. Mill seems to agree with interpreting cognitive defects as a problem of character to be corrected by education. Much like Locke, he places a large portion of human reasoning under the aegis of probability – though in him the term is inference, noting that much of what we observe in actuality contains assumed inferences. In other words, what we assume to be strong, undeniable connections between facts are most of the time associations that we make thanks to past inferences, which have so far not failed us, but do not form a strong bond as we might think[2]. This probabilistic nature of human reasoning, combined with the pliability of human character leads Mill to speculate about a science to come: *ethology*, understood as "the systemic science treating the topic of how upbringing and environment effect the formation of individuals"[3]. Mill believes ethology could "allow us to determine conditions most conducive or damaging to the production of characters we think desirable". In other words, for Mill, human character can be changed in a methodical manner, and education once again plays an important role for this endeavour. It could be said that Mill's *Autobiography,* which tells the tale of "an education which was unusual and remarkable", is Mill's attempt at auto-ethology, to understand the ways in which his upbringing and instruction helped him overcome the cognitive defects he saw in others of his time.

Mill's solution to cognitive defects also presents a political dimension: while Mill maintains his strong conviction for the pliability of human nature and the capacity for individuals to change their character to achieve higher wisdom and intellectual capacity, he

[1] John Stuart Mill, *Civilization*, XVIII: 145

[2] *idem*, *System*, VIII: 641–2

[3] *ibid*, VIII: 861

believes applying this principle in real-life runs into practical problems. Simply put, Mill believes that for the majority of human population, dedicating their time to increasing their wisdom is simply not feasible, not when they need to spend the majority of that time pursuing ways of subsisting[1]. Thus, formal education takes on the task not just of educating people to improve them, but also to help them know *who knows better*. Locke imagines individuals who receive such an education as an "enlightened public" who are "able to discern who are those that know them better[2]. Since Mill clearly establishes a hierarchy of *experts* who have better rid themselves of cognitive defects and others who should be made to recognise this fact, his political system takes this hierarchy into account. This ties into Mill's general view that the form of government in a nation decides that nation's character. In other words, for Mill, government is also an *educational* tool, and representative government is valued for the character traits it fosters as well (in this case, deference to those who know better). The level of education an individual receives should decide, in Mill's opinion, the number of votes they should cast, thus giving a higher importance to the opinion of those who are "better and wiser"[34]. Mill's analysis also seems to focus on cognitive defects as a fault in character, and thus a fault in choosing the right ends.

A final comment should be made on the strong connection Mill makes between the form of government and the character (and consequently, cognitive defects) of citizens. As Glyn Morgan puts it, this makes Mill a "social holist", who "cannot generally expect the character of individuals in a society to depart too far from the broader national or societal culture"[5]. As a result, he deems only *certain* forms of government to be civilised and encouraging liberty, and ends up defending "a form of Western imperialism" for those who would fail this standard, and would need to be governed by others until they were decided to be educated enough (again, by

[1] *idem, Spirit of the Age,* XXII: 241

[2] *idem, Inaugural Address,* XXI: 223

[3] *idem, Considerations,* XIX: 474

[4] *ibid,* XIX: 478

[5] Glyn Morgan, *The Mode and Limits of Mill's Toleration*

the imperial governors). The rhetoric of cognitive defects thus obtains an imperialist character when it is used to reduce "savage" peoples to the same status of irrationality as children[1].

3) Bentham's Diminishing Marginal Utility: Means-Based Defects

It should become apparent that early modern authors gave a larger consideration to the defects of ends compared to the defects of means. There are, however, some notable examples of means-based defects that were studied by these authors, and who offered means-based solutions for them. The example we will consider is Bentham's diminishing marginal utility.

The concept of diminishing marginal utility (in other words, the idea that additional increments of pleasure became less and less important for an individual beyond a certain point), as described by Bentham, naturally fits into our conception of a means-based cognitive defect. The idea of an action becoming ineffective *as an instrument* regardless of its aims is discussed by Bentham to great effect: Since losing a portion of his wealth will hurt a rich man less than the pleasure such wealth would bring to a poor man, Bentham proposes distributive policies as a way to increase public interest[2]. Similarly, Bentham puts greater care to the protection of existing property and the alleviation of suffering due to the people's perception of such losses being disproportionately higher than any additional wealth[3]. Thus we see that Bentham considers the efficiency of these means as means themselves, in light of people's cognitive defects in taking these losses and gains into consideration, instead of perceiving it as an arithmetic progression.

The means and ends-based cognitive analysis of early modern philosophers thus seems to offer a robust theory of action, with separate proposals to "repair" irrationalities in each hub. Following early modern thinking, fixes to our cognitive defects change the agent's disposition to prefer a goal or means that they would deem as better had their cognitive defects not diverted them. Contemporary account of cognitive defects, in turn, will offer a much more complex and interconnected account.

---

[1] Mehta, Uday Singh, *Liberalism and Empire*, p. 102

[2] Bentham, *Theory of Legislation*, I, 103–9

[3] *Idem*, *Jeremy Bentham's Economic Writings*, III, 324, 342

B - Contemporary Defects

In many respects, the contemporary understanding of cognitive defects builds up on our early modern understanding of rational action. We will see that the empirical and utilitarian theories of rational action become the foundation of a more comprehensive rational choice theory, with a far more complex account of cognitive defects within. This chapter will first offer a quick account of Jon Elster's Rational Choice Model as a complete normative theory of decision-making, from which behavioural sciences have observed predictable deviations. These deviations will be studied not only as problems to be solved, but tools to be used in the next chapter.

1) Jon Elster's Rational Choice Theory: Beyond Means and Ends

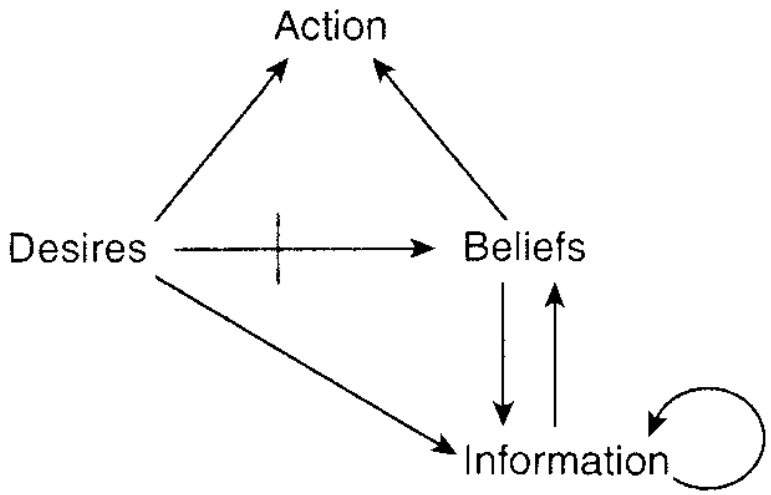
There exist many accounts of choice theory in contemporary theory, stemming not only from psychology (as would be expected by Mill's ethological predictions) but also economy (with the express purpose of explaining decision making as a calculus of maximisation) and sociology (with the aim of explaining individual action through external factors). When it comes to offering a meaningful account of cognitive defects in contemporary political theory, however, not all accounts are equal. First of all, a useful model for our purposes should be able to both judge decisions as rational or irrational. This needn't be able to explain any and all real-life human behaviour, but it should provide a satisfactory definition of rational action so that the course of action a rational agent *would* pick is sufficiently charted. Furthermore, for the cases where an agent doesn't follow the rational course of action, the model should be able to show the particular way in which the agent deviates from rational behaviour, so that we may develop the patterns of irrationality that will form our cognitive defects.

The particular way in which Norwegian social and political theorist Jon Elster understands Rational Choice Theory fulfils all of our expectations. Elster used his particular interpretation of this theory to apply it to many different topics in his writing, though he deemed it less and less as an explanatory and more of a normative model as his writing progressed[1]. The main reason for Elster's progressive distrust of rational choice theory as an explanatory

[1] Jon Elster, *Solomonic Judgements,* p. 1: "As argued below, rational-choice theory is first and foremost a normative theory and only secondarily an explanatory approach."

model seems to be its shortcomings: in other words, the considerable number of cases where humans seem to deviate from it in a predictable manner. While we will discuss the exact nature of these deviations below, so far this model seems very appropriate for our purpose: it offers a model of rational action, with well-documented ways in which observed behaviour strays from it. Let us now see the exact nature of Elster's interpretation of rational choice theory.

Elster defines the relationship between the components of rational choice theory in a simple diagram:



Following this diagram, an action is rational if it is "the best way to satisfy the agent's **desires**, given its **beliefs**". Each part of this diagram deserves to be examined more closely, with an examination of its relation to every other part.

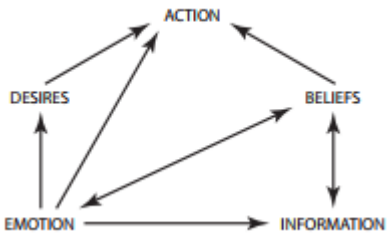
The **desires** and the **motivations** of the agent are one of the major components of action. They are the primary drive for action, and a rational action should offer the best way to bring them about. In cases of irrationality, desires can shape **beliefs** and the **information** the agent collects (Elster gives the example of the enamoured person who becomes convinced that the object of their desires loves them back).

The **beliefs** of the agent form the second component of action. They are the frame through which the agent places their action in the context of the larger world. In other words, they're what the agent believes to be true about their action. For an action to be rational, its efficiency must be justified by the given beliefs. Beliefs are formed by the **information** the agent gathers before deciding on their course of action. In cases of irrationality, beliefs might

be shaped by **desires** (same example as above), or there may cases where information received may be distorted in predictable ways – what we can call cold sources of irrationality, or **cognitive biases** (which we will go into in great detail later).

The **information** the agent collects to inform their desires and beliefs is anything exterior to them that would factor into making a decision about the related topic. The agent is assumed to dedicate the appropriate time to collect the appropriate information before making a rational decision (appropriate being the amount of information necessary given the importance of the decision – an important decision with great impacts would require more information while a simple decision with little importance doesn't require much mulling over). An irrationality then can occur if the agent spends *cannot* decide on the appropriate time to dedicate for a given decision – mostly by not giving the decision enough consideration (due to a sense of urgency created by intense emotions).

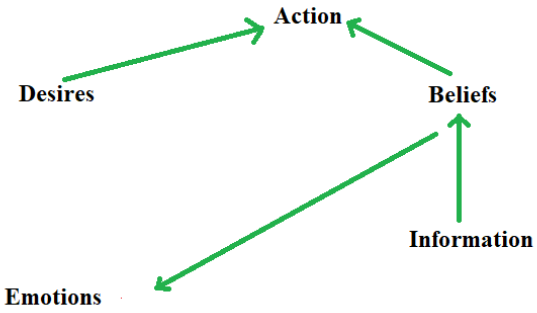
Elster reuses this diagram for the Rational Choice Model in several of works, albeit with slight differences. Some iterations also incorporate emotions as a separate factor of irrationality in the following manner[1]:



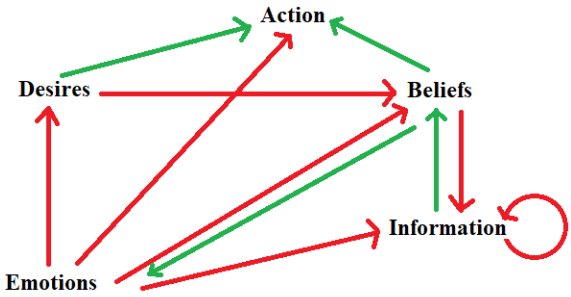
This figure aims to show that emotions can affect all corners of the diagram (they are shaped by beliefs, shape beliefs and desires in turn, can cause an urgency to disrupt information collection and thus cause premature action).

To reiterate the diagram in a simplified manner, the case of a rational decision with no deviations can be drawn in the following way:

[1] *Idem*, *Reason and Rationality*, p. 41



Whereas the diagram with all possible irrationalities changes in the following manner:



An initial reaction to the complexity with which rational choice is presented in this model might be befuddlement, but the way the model functions is in reality not too far from the ends-means model. One could, without causing much controversy, say that ends as a set can be considered as within an agent's desires or motivations, and the means they set for themselves, inasmuch as they serve as the frame through which the ends are obtained, function like the beliefs of an agent. Thus, any argument for rationality that was examined through the ends of means-ends rationality can be translated without much difficulty to Elsterian Rational Choice Model.

Note that this does not mean there is a perfect one-to-one correspondence between the two models. This is mainly due to the Elsterian model covering a larger ground than the means-ends model. The beliefs of an individual to which their action must conform in order to be rational, includes a much larger set than simply the means they have set themselves. It could, for example, include facts about themselves that they believe to be true, which wouldn't normally be considered within means for the purposes of rationality. To return to the example of protesters, the *belief* that a policy is unjust is given consideration in the Elsterian model; in

the means/ends model we sketched before, there would be nowhere to place this proposition. In a similar vein, desires and motivations cover a much larger ground than just ends that an individual has set in relation to their action. Elster gives passions as an example of motivations one might place in this model, and madness or intoxication as particular examples[1]. Neither would fit within the definition of an end, but they are covered by the Elsterian Rational Choice Model as motivations.

We can see that Elster's Rational Choice Model offers a fairly robust account of rational agents, and we can also see how each variable would lead to a specific deviation from rationality due to its characteristics. For the sake of completeness, we should also note that Elster explains another way in which his rational choice model fails to explain the actions of real-life agents: indeterminate situations. By indeterminate situations, Elster means choices made by agents where rational choice theory is *unable* to make a unique prediction as to how the agents will act[2]. Elster demonstrates cases of indeterminacy through hypothetical scenarios where the conditions of rationality for desires and beliefs can't be met at the same time (e.g. Soviet dissidents who hold onto an irrational belief if they desire to resist, but lose a rational basis for their desires if their desires are malleable to external oppression – i.e. are not theirs), or where the amount of uncertainty makes any meaningful decision impossible, or where fulfilling rationality requires an infinite recursion (as is the case of gathering information for some decisions, where the agent has no way of knowing how much information is necessary, and this uncertainty leads to an infinite recursion of irrational decisions). While cases of indeterminacy play a central role in Elster's criticism of rational choice theory, they are not problems of cognitive defects so much as paradoxes within the axioms of rational choice theory. As such, we will not be paying them any further attention.

With this sufficiently thorough examination of Elster's rational choice model as a contemporary account capable of considering cognitive defects, we can now consider what these cognitive defects precisely are. There, (much like Elster) we will defer to the works of cognitive psychologists and behavioural economists.

[1] *ibid*, p. 46

[2] Conversely, cases of irrationality pose a problem because they cause the model to be *inadequate* – the model is able to make a prediction, but its prediction does not come to pass.

2) Cognitive Biases, Defects as Studied by Psychology and Economics

Much like cognitive defects as described by early modern philosophers, cognitive defects as studied by contemporary sciences cover multiple areas where rationality may break down. Just as early modern philosophers' accounts of cognitive defects could be classified on defects of means or ends, most cognitive defects as described by cognitive psychology and behavioural economics can be classified under defects of desires or defects of motivations. There are, however, certain cases where the defects occur due to the interaction *between* desires and beliefs (on which we have touched upon very briefly during our initial exposé), as well as some defects that are in fact tied to cases of indeterminacy.

a) "Hot" Cognitive Biases: Rational Decision and Passions

The first cognitive bias that we may consider is one that is caused by "passions" under Elster's definitions, or what cognitive scientists call "hot" cognitive defects. While Elster briefly blames passions as the reason for weakness of will (where the individual is unable to make a rational decision due to one passion overpowering the totality of the agent's desires), cognitive sciences offer a fairly concrete case of how they work. Experiments conducted by Dan Ariely in the University of Berkeley asked students whether they would follow with certain morally questionable actions[1]. They answered these questions in a calm, non-aroused state as well as in an aroused state. The answers of the same students varied by shocking amounts depending on their level of arousal[2]. Thus, cognitive sciences seem to give empirical proof of not only passions leading to irrational decisions, but also agents being unable to guess *how much* passions can affect their decision-making skills.

Another example of hot cognitive defects relates to the problems of weakness of will. While the topic has been studied in many different forms since Ancient Greece (where it was known under the name of Akrasia), cognitive sciences offer an account of immediate "hot" passions defeating "cool" long-term planning. Behavioural economists seem to explain weakness of will as a result of decisions being taken in a calm state where the time to act on the

[1] Dan Ariely, *Predictably Irrational*, p. 89-108.

[2] As an example, the students were 25% more likely to predict they would forego condoms while aroused compared to their predictions in a calm state, while their predictions on what activities they would prefer to engage in could vary as high as 229% between nonaroused and aroused states.

decision is away, and the agent is able to weigh what they think they should do against other, more gratifying options in a rational manner. When it comes to act on the previous decision, however, the emotions provoked by more gratifying options become more desirable to the agent than what they had initially decided: the result is a scholar whose article should have been finished weeks ago, for which the act of writing is constantly delayed.

b) Beliefs Shaping Desires: Framing, Conformity and Others

While hot cognitive biases may be the product of passions shaping our desires (a case foreseen by Elster in the explanations of his diagram), studies seem to show that certain cognitive biases indicate an irrationality with a mechanism he hadn't elaborated: beliefs directly affecting desires and motivations. While Elster had accounted for desires affecting beliefs (what we desire shaping our beliefs in an irrational manner), the opposite is also true in some cases.

Which specific cases? For our purposes, three specific biases stand out: *framing effect*, *placebo effect* and *social conformity/need for uniqueness*.

Framing effect refers to the predictable manner in which people's decisions change depending on how their options are presented: an action with risk is preferred when presented in a negative frame, and avoided in a positive frame[1]. This is explained by behavioural economists as a result of the higher importance individuals place in losses in prospect theory, meaning that presenting the same option as either positive or negative changes the way it is appraised, and leads to a different outcome. When placed in the mechanisms of Elsterian rational choice, this can be explained as a belief (an option presented is believed to be a gain or loss, depending on the frame) creating a different desire (a stronger desire if the frame presents the situation according to losses, or weaker in the reverse situation). Thus, our beliefs seem to affect the preferences that form within as desires in an irrational manner.

Placebo effect is another well-documented bias that works on a similar principle. Professor Irving Kirsch hypothesised that the belief that a patient is receiving treatment can create subjective changes – which means that the way the agent appraises the situation is modified by their beliefs[2]. Their appraisal of the situation can be considered as their desires and

[1] Sunstein & Thaler, *Nudge*, p. 36-37.

[2] Irving Kirsch, *Response Expectancy as a Determinant of Experience and Behavior*.

motivations in our Elsterian diagram: the agent feeling that their headache has disappeared is, by a certain definition, the end that the agent is pursuing. As we had said that ends map onto desires and motivations in the Elsterian model, the response expectancies of placebos can be considered a desire. Thus, we end up with another scenario where the beliefs agents have about how their action will result will create the very end they hope to achieve, thus their beliefs will be actively shaping their desires.

Lastly, the two different biases of conformity effect and need for uniqueness, which operate on the same principles to produce opposite results, can be seen as cases where beliefs shape desires. Social conformity refers to cases where individuals' preferences (i.e. their desires) conform with what others around them want – thus individuals can desire something they initially hadn't (or wouldn't if others around them hadn't desired it) to conform with the rest of their social group. Conversely, individuals might form preferences for options not expressed by others if they perceive a social expectation to have a different preference (Dan Ariely gives the example of ordering food: people in the US seem to statistically prefer ordering items not yet ordered by others when ordering in groups, thus shifting their preferences in opposition to others). Both cases show beliefs about what the agent *should* choose according to social norms (whether those norms demand conformity with the larger group or divergence from it) modifying the desire the agent would have had their decision been made in a vacuum[1].

All three examples of biases studied by behavioural economists and cognitive psychologists thus show that beliefs (as we classify them in the Elsterian model) can have an irrational effect on the desires the agent forms.

c) Preference for Status Quo: Biases in Desires

The biases explained so far have dealt with the agent's desires being irrationally changed by a different mechanism in the Elsterian model (emotions or beliefs). However, cognitive sciences suggest the existence of an inherent preference in desire formation that favours some situations more than others. This inherent preference can be summed up as an inertia that works against any situation that is perceived as different from the current course of affairs: a bias for the status quo.

[1] Ariely, *Predictably Irrational*, p. 235-236.

Status quo bias means that individuals prefer the current state of affairs for no reason beyond the fact that it is the way things currently are: thus, this bias differs from cases of conformity or passions where beliefs and emotions change the desire that the agent forms: here, the desire formation is biased without any reference to a different mechanism[1]. One explanation for the status quo bias might be a conservation of mental effort: Making a decision requires some mental effort to be exerted, while continuing the already existing course of action would not require a new decision. As a result, changes to any perceived status quo might be seen as an extra effort, meaning that the formation of desires inherently prefers not creating extra effort if possible.

As we will see in later chapters, status quo bias allows certain courses of action to be preferred above others, which can both be seen as a cognitive defect, and can be used to devise solutions *for* other defects.

d) Heuristics and Probability Calculation: Defects in Belief Formation

So far we have focused on the many different ways desires in a rational agent may become irrational: either through the intervention of other mechanisms in the formation of a desire, or through inherent biases in desire formation for certain states above others. For the formation of beliefs, we shall see that humans' "cold" cognitive biases cause them to appraise their beliefs in irrational ways.

Considerable amount of literature in contemporary studies of cognitive defects is dedicated to the study of cold biases. These biases tend to present themselves as predictable deviations from correct probability calculations: in other words, they seem to relate to our inability to assess risks correctly. When making probability calculations in our heads while making quick decisions, we seem to use what cognitive scientists have dubbed "heuristics"[2]: mental shortcuts that do not run full calculations, but tend to give reliable results nevertheless. The reliability of these results becomes demonstrably false in specific situations, however. Thus, these heuristics cause us to form incorrect beliefs, which cause us to make irrational

[1] Sunstein & Thaler, *Nudge*, p. 34-35.

[2] *Ibid*, p. 23.

decisions that a fully-rational agent (in other words, an agent who would run the full probability calculations instead of using heuristics) would not. As far as our Elsterian modal is concerned, heuristics and the biases they create are cognitive defects within belief formation (in the same way that status quo bias is a defect within desire formation).

As previously stated, a large body of works exists on many different heuristics humans seem to use when calculating probability. While listing all of them would be pointless, cognitive scientists seem to focus on some specific heuristics, and these recurring heuristics will be the ones Nudges primarily aim to solve: anchoring, representativeness bias and loss aversion.

Anchoring, in the case of probability calculation, is also known as the adjustment heuristic. It refers to cases where agents take an implicitly suggested reference point (which thus acts as an "anchor") and take a result around that point while taking shortcuts in calculation. As a result, the numbers people calculate can be modified simply by anchoring them to a different number before they are asked to make a quick calculation. This bias seems to function even in cases of subliminal anchoring, or when anchoring is done without any relation to the calculation that will later be demanded. Anchoring thus seems to form a powerful bias in belief formation.

Representativeness bias also skews probability calculations in a considerable manner. This heuristic can be explained as the tendency to give a higher probability to two attributes (thus increasing their probability together for the agent) if the agent believes them to be similar or related. The common example given is the belief that a person is more likely to have attribute A *and* attribute B than having *only* attribute A if A and B together seem to fit their description better. Simple mathematics makes it impossible for A *and* B to be more probable than only A, but the belief that A and B is *more representative* than only A for the attributes of the person causes us to form a wrong belief. The effects of representativeness are particularly important for political decision-making: wrong probability calculations due to this bias can cause political decisions to be unfairly biased toward certain groups (because they are associated with negative qualities) or attribute likely outcomes of a given policy in a completely unrealistic manner.

Lastly, loss aversion in which individuals are more affected by the loss of something than its gain (and in a similar vein, endowment effect where the possession of an object makes it more valuable than it would be deemed otherwise) is a fairly well-studied heuristic. In fact,

Jeremy Bentham's concept of diminishing marginal utility seems to be an early modern study of the same concept. This is deemed to be a bias since not only is the same value is treated differently if it is a loss or a gain, the combination of this heuristic with the framing effect means that the same course of action can be presented either as a loss or a gain – and the agent would make a different choice accordingly, thus rendering their decision-making process irrational.

C - Cognitive Defects: New versus Old

What, if anything, can be said to differ in the way early modern philosophers have treated cognitive defects, and the way in which cognitive scientists currently treat them? Are there meaningful changes in our new bias-based understanding that would invalidate previous accounts, especially in regards to the solutions they had proposed?

A quick glance reveals a key difference in the reason behind cognitive defects. Early modern philosophers seem potentially more optimistic in terms of people's desire to be rational, and their ability to become more rational. It's true that not all early modern philosophers shared Descartes' idea that *"Common sense is the most fairly distributed thing in the world"*, in other words, that everyone had more or less equal capacities concerning rationality. Nevertheless, there seems to be a general confidence in the possibility to improve rational faculties, and that failing to adhere to rationality is more often than not a problem of *character*. This hope[1] in improvement and placing the blame on poor character mean that cognitive defects find their solutions through *education* (an exception being Bentham's more pragmatic approach of working around irrationality).

Contemporary approach to cognitive defects, in contrast, sees them as something that's inherent to human mental faculties. The problem isn't that people are educated poorly and thus calculate probabilities incorrectly, it's that we all have an inner *tendency* to jump to wrong calculations (and consequently, to wrong conclusions), which we can control (and even that hardly) through strict avoidance of spur of the moment decisions. Cognitive sciences see cognitive deficiencies as a systemic problem, one that can be causally observed in human behaviour and whose patterns are predictable in everyone. This systemic (and somewhat

[1] Though it's debatable how optimistic Locke was on the possibility of improvement.

fatalistic) approach to our cognitive defects naturally demands different ways of resolving these defects, as we will see in the next chapter.

II - Nudging and Other Solutions to Cognitive Defects New and Old

The previous chapter started with the early modern understanding of cognitive defects, where solutions were based on education and in some cases making defective thinking align with rational choices (such as Bentham's idea of making personal interests align with public interest). Contemporary accounts of cognitive defects, compared to early modern accounts, see them as inherent error in our decision-making, or systemic tendencies that exist in everyone. This approach naturally leads to different solutions for the problems it perceives.

The following chapter studies the particularities of contemporary solutions for modern defects. It is primarily interested in *Nudge* theory by Richard Thaler and Cass Sunstein, who offer a comprehensive set of principles on how to deal with cognitive defects. Nudges, in their myriad forms, can be grouped under three types of solutions depending on how they treat the cognitive defect they aim to solve: they try to either *remove* a defect, *work around* a defect or *pit one defect against another*. We will see that the last family of solutions is an innovation on early modern thinking. A discussion on what this innovation means for liberalism and democracy is the focus of later chapters.

A - Libertarian Paternalism: Principles of Nudging

Just as the particular way in which empiricists and utilitarians conceptualised cognitive defects shaped the solutions they devised for them, the particular way cognitive sciences explain cognitive defects has shaped the solutions that have been devised for them. However, the solutions to new cognitive defects are also marked by a particular design philosophy. This design philosophy is summed up as *libertarian paternalism* by Thaler and Sunstein, whose work on Nudge theory can be seen as the flagship of the approach we're studying.

Libertarian paternalism as a term sounds like an oxymoron at first glance. Libertarianism is commonly associated with the belief that the free market is inherently just. This stems from the belief that people have a "*right* to dispose freely of their goods and services"[1]. The emphasis thus seems to be on freedom of choice and autonomy as supreme

[1] Kymlicka, *Contemporary Political Philosophy*, p. 102-103

goods. Here it is juxtaposed to paternalism, which is commonly defined as the limitation of others' liberties for their own good (as much as such a good can be perceived). There thus seems to be an inherent contradiction in the terms. Thaler and Sunstein, however, insist on the term as their design philosophy for Nudges on multiple occasions[1]. It seems however that their understanding of libertarianism and paternalism is quite different from the way these terms are normally used. Thus, an understanding of libertarian paternalism requires us to adopt Sunstein and Thaler's outlook. Once we have noted the specific definitions they use, we can see if the two terms can somehow make sense together.

Libertarian paternalism is a combination of the two words with very specific definitions in mind. For Sunstein and Thaler, a policy is paternalistic if it *influences* an agent's choices with the aim of making them "better off". Their conception of paternalism thus seems wider than the commonly-held definition (where, as previously stated, paternalism is seen as the *limitation* of autonomy). Furthermore, their use of making an agent "better off" is equally noteworthy. Their paternalism thus hinges on a comparative conception of goodness, and whether this "goodness" is based on happiness, utility, efficiency or some other factor is yet undecided (though as we will see, the full definition can give us an idea). Thaler and Sunstein's conception of paternalism is thus one where the paternalistic figure is assumed to know the superior state the agent should end up in, and where it exerts an undefined amount of influence over the agent so that they end up there. But this is not without problems. In order for this definition to be admissible, the "better off" comparison must work. This means that whatever idea of "goodness" nudges settle on, the paternalistic principle must justify the incremental gain it offers is enough of a justification to allow an intervention. It may not be paternalistically justified to limit a morbidly obese person's eating habits if it makes them lose a tiny amount of weight, but make them miserable in the process. On the other hand, limiting a bus driver to only 8 hours of driving per day could be justified if it shows to prevent major car crashes – i.e. if the increment of change is more meaningful. Thus, when it comes to measuring how "better off" an agent gets with a paternalistic policy, the type of goodness we measure, as well as the increment after which an improvement justifies paternalism both matter. Once we have the full

[1] Thaler & Sunstein, *Libertarian Paternalism*; Sunstein & Thaler, *Libertarian Paternalism is Not an Oxymoron*; Camerer, Issacharoff, Loewenstein, et al. *Regulation for Conservatives: Behavioral Economics and the Case for "Asymmetric Paternalism*, among others.

definition of libertarian paternalism with a clear idea of the superior state we're seeking, we'll need to come back to this question.

When it comes to their definition of libertarianism, we can also find some differences compared to the common conception of the term. The common understanding of libertarianism focuses on *autonomy* as a principle, with a trust in free markets as a result, and Thaler and Sunstein define libertarianism as the defence of "freedom of choice"[1]. By focusing on freedom of choice, having as many choices as possible, and choosing what you *truly* want out of them becomes morally valuable. The more options there are, the freer the individual is, and more preferable is the configuration from a libertarian perspective (at least according to the way Sunstein and Thaler define such a system). This notably makes no distinction on where choices are multiplied. Under libertarianism, having as many ice cream flavours as possible is of equal moral worth as having as many health insurance policies as possible. For now, we will limit ourselves to only noting this property. It will come up again when we treat the compatibility of libertarian paternalism with liberalism.

When combined with these particular definitions, libertarian paternalism starts to make sense: paternalism as *influence* over decisions for the better isn't antagonistic to libertarianism, because it does not reduce the number of choices in a given configuration. Had we taken paternalism in the sense of limitation of liberties, such a view would be incompatible with libertarianism, since the number of choices would be reduced to the ones the paternalistic figure finds in line with the agent's perceived good. When paternalism expands to include cases of influence, methods that don't limit an agent entirely also fall under its purview. As a result, while paternalism by itself still maintains a right to limit the agent, it also becomes open to methods that favour influencing the agent, allowing paternalism to avoid offending (so to speak) libertarianism.

In a similar vein, the understanding of freedom as the freedom to choose allows libertarianism to function without offending paternalism. Under a common definition (where paternalism is *limiting* freedom for the good of the agent), an agent being able to choose an

[1] Thaler and Sunstein make a particular reference to Milton Friedman's *Free to Choose* as the basis of their definition. I leave the reader to decide whether they refer to the book or the public television broadcast.

option deemed worse would run contrary to paternalism's aims. If focus was given to libertarianism as autonomy, as the ability to set one's norms, one would need to guarantee internal freedom, and the two concepts would inevitably be in conflict. But if libertarianism contents itself with maximising the number of choices, it takes into account only external markers of freedom. Thus, paternalism as *influence* runs under libertarianism's radar, since there is no "formal" barrier to the other options, libertarianism comes out unharmed.

Libertarian paternalism starts looking like a plausible combination after all. The peculiarity of this combination allows us to answer some initial questions as well: the question of what "better off" means in particular. If libertarian paternalism assumes people are at their most free when they have as many choices as possible (thus putting the individual's action of choosing at the centre), that agents can be influenced by paternalistic systems (thus implying that individual choice isn't free from external influences), and that individuals' best interests can be known by third parties (thus implying that there is a communicable – if not objective – way of deciding the best course of action for someone), we can arrive at a definition of an agent's welfare that is compatible with the Elsterian Rational Choice Model[1].

The conditions we have just stated above show the compatibility of libertarian paternalism with the Elsterian model: there is an emphasis on the individual's desires (which can be tied to the freedom of choice libertarian paternalism underlines), while admitting that the choice made can be modified by external influences (which allows the existence of paternalism as influence). Furthermore, while the individual's desires and free will play a central role in defining what the desired action is, the Elsterian model gives us objective conditions which can be used to determine whether an action that's chosen is rational given the desires and beliefs of the agent. The Elsterian model thus shows its compatibility with the "better" choices that the paternalistic figure should be able to know. Thus we arrive at libertarian paternalism as a fully-fledged ideology: a design philosophy to help devise solutions to the limitations of the Elsterian Rational Choice Model: the contemporary way of overcoming irrationality will be *influencing* agents to make better choices *according to their desires and beliefs*, while giving them as many choices as possible.

[1] After all, our study of Nudges has the aim of solving problems revealed by this model.

B - Nudges or the Three Methods of Debiasing

What is the practical application of libertarian paternalism when it comes to solving cognitive defects? This chapter will offer a detailed description of Nudges as a policy tool to overcome cognitive defects. Such a description will allow grouping Nudges into three distinct groups depending on how they tackle cognitive defects: *training* agents to undo these defects, *working around* these defects by creating new *incentives* for the "better" options, or (and this method proves to be the novelty of Nudge theory) *using one defect to overcome another* by changing the *choice architecture*.

Nudges as a political policy tool are distinguished from more classical tools of governmental authority (such as mandates, bans, subsidies or fines), according to Sunstein[1], by their *liberty-preserving* approach. This approach is defined by the fact that nudges neither *limit* any given array of options like a mandate or a ban (respecting the libertarian approach), nor add any additional costs to a choice like a subsidy or a fine[2]. Thus, following the libertarian paternalist credo, Nudges applied to any given situation maintain a given array of choices while influencing the agent to a goal they would deem as better had their cognitive defects not diverted them.

In a 2014 essay[3], Sunstein lists what he deems to be the 10 most "important" Nudges: default rules, simplifications, uses of social norms, increases in ease and convenience, disclosure, warnings (which for our purposes can be explained together with disclosure), precommitment strategies, reminders, eliciting implementation intentions and informing on choices. While the list is certainly not exhaustive, going through these different examples should help with finding their common ways of working, which will later help us define the major categories according to the way they deal with cognitive defects:

[1] Sunstein, *Nudging: A Very Short Guide*.

[2] Whether Nudges that make a choice more bothersome (which would add, theoretically, a *cognitive* cost to choosing it) count is a Nudge is a question raised by Sunstein and Thaler themselves. Their answer is a yes *only if the costs are low*, though how low cognitive costs should be calculated is something they do not elaborate.

[3] *Ibid.*

1) Default Rules

Perhaps the most emblematic (and according to Sunstein, the most effective) of Nudges, default rules denote policies where governments (or for that matter, private interests) allow agents to choose between different options with no limitations (e.g. between different saving plans, user interfaces, printing style, etc.) but revert to an option that the *choice architect* (i.e. government or any other body enacting the said policy) deems the best for the agents. This Nudge works by relying on the individuals' status quo bias: since the majority of people will not be bothered to change the default option, any cognitive defects that would lead to inefficient or harmful results are avoided.

2) Simplification

Simplification as a Nudge refers to policies that present existing choices or government programmes in simpler terms, to reduce the cognitive cost of considering the subject and thus making a decision. An example Sunstein and Thaler discuss in their book as a possible simplification is a RECAP system for Plan D prescription drugs under Medicare in the US, whereby citizens covered by the programme would be sent a simplified version of the 46 different prescription options among which they are supposed to make a choice[1]. Such simplifications in Thaler and Sunstein's opinion allow those that can't bear the cognitive cost of processing all the necessary information to make a better choice following the simple explanations, while still allowing a detailed look for those so inclined.

3) Use of Social Norms

Using social norms as a Nudge can be summed up as the positive use of social conformity bias. As experiments have shown that people conform more when reminded that a majority makes the socially desired choice, these nudges can influence agents just by highlighting the desired behaviour. Of note should be that social norms can be used even when they do not truly exist: research suggests Nudges towards more desired behaviours that are currently rarely done can still be effective if the social norms are framed in the right way (effectively giving the impression that the desired behaviour exists in the agent's perception)

[1] Thaler & Sunstein, *Nudge*, p. 159-174.

or if focus is given to what the majority *thinks* should be done even if the same majority deviates from it. Concrete examples given by Sunstein include "90% of people in Ireland believe people should pay their taxes on time", which reduced payment delays even though the majority normally fails to follow this norm.

4) Increasing Ease and Convenience

Another type of Nudges can be explained as "making better choices easier". While Sunstein's phrasing makes the method sound similar to simplification, simplification deals with the reduction of complexity in a choice. It is about the presentation of choices, where *every* choice is simplified, stripped of jargon, and where their notable results, costs and benefits are made clear. Increasing ease and convenience, on the other hand, is about making *better* choices easier *to choose*: it's about the reduction of cognitive costs of only *certain* choices that the choice architect has deemed more preferable. Either by making these options more visible, or making their choice demand less cognitive effort, agents' preference of better options can be increased in a statistically significant manner. In such a Nudge, an agent's desire for immediate gratification (which would normally make him prefer a quicker solution and thus lead to a weakness of will) is nullified by making the rational options equally gratifying in the short run. An example of making an option easier would be putting healthier foods on eye level on supermarkets. By making the better option comparatively easier to get (while maintaining the other options), this method preserves freedom of choice while statistically improving preference for "better" choices.

This policy evidently implies its opposite, whereby choice architects can make undesirable options more difficult to choose, thus reducing the preference for them. Such an option would still remain within the bounds of libertarian paternalism for Sunstein and Thaler if the added cost is "low". Following the supermarket example, putting unhealthy foods in a spot requiring more effort to get to, or making cancellation of gym subscriptions more difficult, or having cool-off periods for certain undesirable actions (joining a gambling club, buying firearms or even filing a motion of no confidence) would achieve the desired effect as Nudges.

5) Disclosure and Warnings

The practice of announcing relevant information that normally wouldn't be reachable without conscious effort takes two different forms of a nudge for Sunstein: disclosing information (where information is used in a more neutral sense) and attaching warnings for undesired effects (where information is specifically about the dangers of an irrational behaviour the Nudge wants to avoid). Examples include anything from announcing the environmental costs to energy consumption, open government initiatives allowing accountability to warnings on unhealthy food, cigarettes or alcohol.

These nudges run under the assumption that people would be making a rational choice and avoid falling to their cognitive defects if they received sufficient information to activate a more rational decision-making pattern. In this case, simply giving relevant information is enough to undo these cognitive defects.

Disclosures and warnings are possibly the closest among contemporary nudges to early modern solutions to cognitive defects. They can be seen as a concrete interpretation of "education" that Mill and Locke proposed, and this similarity raises questions about their design philosophy. Can we truly call these techniques paternalistic if all they do is raise an information that is supposed to be neutral (if it were not, it would fall under conformity nudges or outright propaganda)? After all, what's "paternalistic" about a stop sign? Here, the particular definition of paternalism Sunstein and Thaler employ becomes important: for them, if a policy influences behaviour (i.e. changes behaviour in a statistically significant way) towards a "better" option, it is paternalistic. Thus, even information panels and neutral remarks can have a paternalistic quality.

Furthermore, one may adopt a Foucaldian point of view whereby everything is political[1]. Under such a view, even an act as simple as bringing a certain quality of smoking forward (and consequently hiding other qualities) has a certain political aim, and cannot be neutral. If no neutral act is possible, even acts of information gain a "biopolitical" aspect and thus become paternalistic. This line of reasoning, however, is beyond the scope of this dissertation.

[1] Foucault, *History of Sexuality*, p. 189.

6) Precommitment Strategies

Sunstein notes campaigns where people commit to a certain action (preferably at a specific time) allow them to perform that action better compared to a decision only at the time of acting. This Nudge aligns perfectly with the Elsterian model of action, as precommitment is thoroughly studied by Elster himself[1]. The mechanism can be explained in the Elsterian model as a rational choice achieved in the present time binding itself against any of the potential deviations in the future. It thus serves as a tool against numerous cognitive defects (effects of passions, time-inconsistency when calculating utility, etc).

7) Reminders

Reminders come in many different forms: reminders to make a choice or providing information that could be relevant to a choice at a critical time are the most common. In these cases, reminders to vote, reminders that a bill is becoming overdue or even Facebook's reminder about an event an hour before its designated start all count as reminders. These Nudges seem similar to precommitment at first glance due to the interplay of different times in decision-making (a reminder set at an earlier date, which aims to overcome a cognitive defect at the moment of decision), but unlike pre-commitment, the moment a reminder is set is not a commitment to a certain course of action. In this sense, reminders as Nudges are lighter in the influence they exert, simply showing relevant information but leaving the decision-making process at the time of decision-making intact. Another way of seeing them is a form of disclosure with a specific time set.

8) Eliciting Implementation Intentions

Another form of Nudging suggested by Sunstein is simply asking people what choice they intend to make in a "cold" (in terms of influence from passions) environment, i.e. before the actual moment of action. Sunstein claims that simply asking parents their future intentions on vaccinating their children, or asking people whether they intend to vote in a survey, makes them more likely to follow through with said actions. Much like precommitment, these Nudges work by giving the agents a chance to consider their choice in a past point where cognitive

[1] Elster, *Uysses Unbound*.

defects are less pressing (though unlike precommitment, no commitment is necessary – simply asking the question is enough). The memory of the agent considering what they would do seems to influence them to take that same decision in a "hot" environment, thus avoiding any cognitive defects that could lead them to worse (as Nudge theory understands it) options.

9) Informing on Choices

The last example of Nudges Sunstein gives relate to information on the effects of past choices for an agent. Sunstein argues that individual agents are often unaware of the full scale of the results their actions entails (i.e. their beliefs are not based on the full extent of relevant information in the Elsterian model), and that public and private authorities with this knowledge sharing this information could be sufficient to prompt a change in behaviour towards better courses of action. Examples may include details about credit card use (allowing a user to notice where they overspend and on what kind of purchases) or websites that inform a user about the environmental impact of their lifestyle[1]. In effect, these Nudges can be considered as a subset of disclosing relevant information, but unlike the previous examples, the information in question relates directly to the agent's past choices.

C - Grouping Together: Nudges by Type

The concrete examples that we have given that were proposed by Sunstein himself show some patterns. Most of the proposed Nudge policies work in a similar manner, applying the same idea to different situations, timeframes or actions. These similarities allow us to define three specific ways in which Nudges interact with the cognitive defects they aim to solve. Understanding the nature of these ways will allow us to approach Nudging in the context of liberal democracies with more nuance later on.

[1] An interesting example is https://www.footprintcalculator.org/, a website that asks users several questions about their habits in eating, travel, energy consumption and so on, and gives a very concrete result – how many Earths would be needed to sustain such a lifestyle if everyone lived in the same manner – in the end. Public bodies could theoretically make similar calculations known without the need of a prior questionnaire.

The three groups we will propose align with three approaches to debiasing that cognitive scientists have identified[1]. Cognitive science categorises these approaches to debiasing as *training*, *changing incentives* and *changes in the presentation and elicitation of decisions*[2]. These three groups are sufficiently distinct to place all Nudge examples given by Sunstein to place solely in one category or another. Nudges that only inform the agent before a decision (disclosing, warnings, informing on choices, reminders, etc.) can be seen as different forms of *trainings*, while Nudges that make better choices equally appealing can be seen as creating *incentives* (increasing ease and convenience, for example), while a great deal of other Nudges concern themselves with the way decisions are presented (default rules, use of social norms).

The correlation is not 100% foolproof, however. Some Nudges don't seem to fall into either category (for example, precommitment can't be said to train the user, neither does it offer a counter-incentive, and nor is it about presenting a decision – rather, it is about making a committing decision at a specific point). Furthermore, while this classification is useful for discussing debiasing methodologies (and serves as a good starting point for us), it does not describe how each Nudge treats a cognitive defect in a way that would be useful for our political analysis. Presented as training, incentives and choice architecture, these solutions can all be politically neutral.

Much like how a means-ends distinction of rationality in early modern philosophy proved to be a useful starting point when relating to the Elsterian Rational Choice Model, the classification offered by cognitive scientists can help sharpen our intuition on what sets these different methods apart. But the core distinction we will focus on is the following: how these Nudges *treat* cognitive defects.

What I mean by treating cognitive defects differently is asking whether a Nudge aims to *remove* a cognitive defect from the decision making process, whether it aims to *work around* said defect and act as a counterweight, or whether it aims to *use* another cognitive defect to solve this one, effectively *instrumentalising* a cognitive defect as the solution to another.

---

[1] Morewedge, Yoon, Scopelliti, et al, *"Debiasing Decisions: Improved Decision Making With a Single Training Intervention"*

[2] While some behavioural scientists calls the last group *nudges*, for us Nudge is a term covering all three solutions, as solutions that would fall into first and second groups such as creating new incentives or disclosing information are also considered to be Nudges as per Sunstein's definition.

Such a distinction makes no change to where the Nudges we were able to group initially stand. All Nudges that aim to *train* an agent hope to minimise the effect of a cognitive defect entirely, thus they stay firmly in *removing the effects of defects* camp as a solution. All Nudges that try to create incentives do not tackle the defect itself, but strengthen a different option so that the decision making process will *work around* said defect. Nudges that change choice architecture are similar to workarounds in that they do not tackle a defect itself, but while a workaround makes other choices more attractive through more rational incentives (monetary gain, ease, etc.), changes to a decision's presentation employ explicitly other cognitive defects: A default becomes more preferable *because of* our status quo bias. Saying "90% of people pay their taxes on time" is effective not because of its moral weight (moral concerns are already considered without such a prompt), but because seeing such a high majority act in a certain direction (whether they actually do or not) activates our conformity bias and pushes us to act similarly.

It could be said that Nudge Theory's biggest addition to the question of how we resolve cognitive defects is this third approach of pitting one cognitive defect against another. Early modern philosophers' solutions to cognitive defects can be grouped into the first two camps with relative ease: the insistence of Locke and Mill on education to improve people's reason and teach them rational thinking is, practically, debiasing trainings expressed in Enlightenment-era terms. Similarly, the assumption that people were self-interested, and that their self-interest should be made to align with public interest through new incentives is not too different from the idea of strengthening more rational choices with new incentives.

But no early modern thinker would go as far as to say a cognitive defect that they have discovered should be used to solve another defect. The closest analogue we can find in pre-contemporary political theory comes from Plato: the tale of the *noble lie* that would be told in the ideal republic. In *the Republic*, Plato suggests that the three classes in the Republic would be given a false mythology (people having essences of *gold*, *silver* or *brass* depending on their social standing, which the through examinations reveal). He justifies this false foundation as a *noble lie*, whose acceptance "would have a good effect [on the citizens], making them more inclined to care for the state and one another."[1] The institution of such a noble lie can be seen

[1] Plato, *Book 3*, 415c–d

as a Nudge that uses the conformity bias, even when the social norm on which conformity is based doesn't actually exist. However, this Platonic conception of noble lie in service of politics is not really used again by any pre-modern or early modern philosopher. This avoidance can be explained through a simple reason: the purpose of the state (and consequently, the purpose of the citizens) has changed from Plato to modern era. For Plato, *the Republic*'s initial premise is finding the objective definition of justice. As a result, the citizens of the ideal Republic exist first and foremost as parts of a machine that seeks to achieve justice. Their value comes from fulfilling this purpose, with everything else coming secondary. The three early modern philosophers we studied were all either proto-liberals and outright liberals. Their conception of the State is one that *serves* the citizens, either as a contract that protects against the state of nature, or as a government whose sole purpose is the maximisation of utility and happiness. Under such a definition, citizens are not seen as cogs of a machine, and thus using their weaknesses against them is not even considered. Thus, nudges remain the first in-depth application of cognitive defects *against* other cognitive defects in a modern point of view.

Thus, our descriptive effort in understanding how cognitive defects are analysed and treated arrives at its conclusion. Equipped with an account of both how early modern philosophy classified cognitive defects and attempted to solve them, as well as how contemporary theory (and cognitive sciences in particular) understands and solves these problems now, we can appreciate the true scope of Nudging as a political tool. While all three approaches to Nudging seem sensible in a vacuum as methods to treat cognitive defects, the ideological question of their political use opens a whole new dimension worthy of discussion. However, it is a bit concerning if the closest pre-modern analogue to nudging is the Platonic noble lie. It shows that the ideological underpinnings of liberalism have never considered such methods before, and perhaps for a good reason.

III - Liberalism vs. Libertarian Paternalism: Problems of Nudging as a Principle

The following two chapters discuss the legitimacy of using Nudges from a political perspective. As the Definitions section in the Introduction have outlined, this dissertation limits its political consideration to a liberal democratic perspective. This means that the primacy of individuals' freedom to form their own preferences, and the organisation of the body politic so that it allows the participation of every citizen in some form, are both assumed. Given our present societies, neither of these assumptions should be shocking, and thus no further effort is made to justify them.

While these core assumptions go unchallenged in this dissertation, their more detailed expressions inevitably clash with the political implications of *Nudge Theory*. These problems can be roughly grouped into two groups: the theoretical contradictions within Nudge Theory (and its underlying ideology of libertarian paternalism) revealed through the lens of liberalism, and the challenges to Democracy posed by the existence of Nudging as a political tool. This chapter will thus approach Nudge Theory and libertarian paternalism with a critical take, using the core tenets of liberalism (with which libertarian paternalism should theoretically be compatible) as its guiding principles.

Several criticisms from the liberal point of view can be raised against Nudge theory and libertarian paternalism. Sunstein and Thaler expect these criticisms to concern the paternalistic side of nudging. However, we can also ask whether the libertarian principles of nudging are compatible with the way liberalism expects agents to be treated. This chapter examines and argues for these fundamental incompatibilities in detail. It also takes into consideration the counter-objections Sunstein and Thaler have raised to similar criticisms. It concludes with the belief that libertarian paternalism fails to offer a robust foundation for political treatment of cognitive defects. The practical questions that an acceptable political understanding of Nudges (especially in regards to democracy) forms the heart of the discussions of the final chapter.

A - Liberalism and the Obligation to Nudge

The most immediate liberal objection to nudging is asking whether we need to nudge at all. A liberal might find the whole idea objectionable since it places the choice architect in a

superior position (thus invalidating the equal consideration everyone should be given under liberalism), and influences the choices of an agent when we may have no right to interfere (say, from a Millian point of view where interference is justified only if the others' security is at risk). Sunstein and Thaler's response to this criticism is fairly decisive: we *have to* nudge, so it's better to nudge according to libertarian paternalism.

The reasoning goes thus: choices happen in an environment that activate certain cognitive biases whether we want them to or not, and it is practically impossible to make choices in an "architecturally-neutral" environment. This structural inevitability means that every policy where agents need to make a choice forces us to design a choice architecture, whether we want to or not, and whether we want to use that architecture for a purpose or not. In order to avoid Nudging people in directions we don't actually want them to take, we have to Nudge them consciously. In which case Nudging according to libertarian paternalist principles is the best option, since it promotes agent welfare (as much as possible according to the agent's own definition) and respects freedom of choice.

If the reasoning's premises and conclusions are correct, we would indeed need to accept Nudging, if begrudgingly. In such a scenario, we would have to accept that libertarian paternalism, while not fully compatible with liberalism in its principles, offers the least damaging architectural philosophy for policy design. Thus, libertarian paternalism would find an uneasy existence within the bounds of liberalism, acting as the "least bad" boundary to the realisation of liberalism to its fullest extent. Such a coexistence would be akin to the practical difficulties of forming a fully representative public opinion in a large-scale democracy.

It is true that some choice environments require a default setting, or similarly social norms are revealed in some situations whether we want them or not. In these cases, a default *has to* be set one way or another. An interesting example for when defaults are inevitable would be Internet browser choices. In 2010, the European Commission investigated Microsoft, the developer of Microsoft Windows (an Operating System then used by 90% of all computers in the EU[1]) and Internet Explorer (an Internet browser then used by ~45% of all computers in the EU[2]), for charges of abusing its dominant position in the market. To settle the issue, the

[1] Statistic retrieved from http://gs.statcounter.com/os-market-share/all/europe/2010

[2] Statistic retrieved from http://gs.statcounter.com/os-market-share/all/europe/2010

Commission obligated Microsoft to display a webpage, BrowserChoice.eu, to allow users to choose their own web browser instead of using Microsoft's default Internet Explorer. The caveat was, of course, that an Internet browser is needed to access BrowserChoice.eu – meaning that no matter what Microsoft did, Internet Explorer was – even if only for several seconds – the default choice.

However, the conclusion that Thaler and Sunstein draw from their admittedly correct premises does not follow. At least not to the extent that they would like it to go. While it is true that some situations inevitably require to Nudge in certain directions, it is by no means evident that libertarian paternalism and its focus on increasing welfare should then be the guiding principle for Nudging. On the contrary, it seems to require a logical leap since the paternalistic premises of libertarian paternalism are not included in core liberal assumptions.

What then, is the liberal conclusion to draw from the obligation to nudge? A liberal might accept that every configuration creates a choice architect of some sort. But they can also say that some choice architectures can be ignored if the structure is liberal enough. Liberalism only demands equal moral worth for each individual and freedom for each. What people do with this equality and freedom, provided that the freedom and equality in question are sufficiently robust, should be completely up to them. The structure of liberal political institutions should be able to hold out on its own without any need for nudging. Take a more controversial example such as drug use: assuming that drug use was to be decriminalised, a libertarian paternalist might concede to such a policy but demand that hard drugs such as heroin be very hard to obtain, that every dose of LSD contain a note on the risks of psychosis upon use, and that marijuana users be informed of their increased food consumption ever since they started smoking. A liberal would simply say that "it is morally indefensible to use the mechanism of the coercive law to attempt to prevent people from engaging in activity that gives them significant pleasure, rarely does them significant harm, and does not harm others any more than many legal activities"[1], and legalise access with no interference provided that external harms such as drug cartel involvement and child neglect due to drug use are dealt with.

[1] Johnathan Wolff, *Ethics and Public Policy: A Philosophical Inquiry*, p. 75

Thus it seems that Sunstein and Thaler's eminent defence against potential criticisms through the obligation to nudge does not hold. With that in mind, we can form our own line of liberal criticism against libertarian paternalism.

B - A Look On Nudges from the Exterior

While it may not be the most evident, the simplest way to start our criticism of libertarian paternalism as an ideological skeleton for nudge theory is with a change of perspective. So far, we kept the Elsterian Rational Choice Model in mind while discussing any politics that would be designed with Nudges: in a way, the internal structure of the agents was visible while examining the way in which certain policies functioned. This undoubtedly allowed us to see how Nudges interact with the internal mechanisms of human decision making and improve them.

However, suspending this transparency could be surprisingly revealing as well. Would imagining every agent as a black box, whose inner processes are not taken into consideration, change the way policies of Nudging would be appraised? When comparing Nudges to lack of Nudges, the general outline would remain the same: a group of individuals would be facing a certain number of options. Without this policy, there would be a certain distribution of individuals towards certain choices. One choice (or a group of choices) would be deemed rationally preferable (following libertarian paternalism's definition, *as defined by the agents themselves*), even if not everyone would be making the rational choice. The only difference between the way we discussed choice so far would be that the exact mechanisms that cause this deviation from rationality would not be explained.

A concrete example would be the following: in a given university cantina, students are given full freedom in terms of the lunches they can choose. In the absence of any policies, 15% prefer salad, 20% prefer fish, 30% prefer chicken with rice, while 35% prefer hamburger and chips every day. In the same student body, 70% agree that a balanced diet is important, and that they should avoid eating too much red meat and high-caloric food. Thus, we see that the rationally undesirable option (hamburger and chips) is nevertheless the most preferred.

If a Nudge is introduced to a system, the choices change in a significant manner, despite all the options keeping the same incentives in terms of economic costs and availability. This change can be observed without referring to the internal mechanisms of the agents.

Following the same cantina example, let us suppose that a Nudge has been placed whereby the cantina prepares salad for every student unless they go out of their way to order something different the morning of the same day. With that change, 60% end up eating salad, 10% eat fish, 10% eat chicken with rice, and the amount of students who choose hamburger and chips drops to 20%.

By all accounts, this is a successful application of a default setting to Nudge students towards healthier options. Students are just as able as before to order the food they want, but eating irrationally attractive food becomes more difficult due to the effort required to order it beforehand. All in all, such a policy is libertarian paternalism at its finest.

However, a curious thing happens when we compare this Nudge to a traditionally paternalistic policy while maintaining our "black box" perspective on the agents' internal mechanisms. Let us take the same university cantina, and assume that the administration outright removes every option except salads from the menu. Assuming that no policy is perfect and that students have other (more costly, both economically and in terms of availability) means of ordering food[1], a new distribution of choices might again be 60% vegan meals, with the rest of the options getting 10% each as students go to a different restaurant, get food delivered, or cook themselves. In other words, from an outsider's perspective, Nudging and traditionally paternalistic policies look the same.

Of course, this similarity between libertarian paternalism and traditional paternalism when the internal decision-making processes are removed is painfully obvious. Furthermore, this similarity by itself is not a proper criticism of the ideology – it assumes that decisions will be influenced. However, it illustrates one point: liberal paternalism, despite its insistence on *respecting* freedom of choice, *affects* the choices that are made. Its whole argument stands on the premise that the effect it has is acceptable for a libertarian (or liberal) worldview while traditional paternalism isn't, because of its "soft" approach: the options are there, it's just that individuals choose them less in favour of "better" options when Nudges are in place. However, as we will see in the next section, "soft" paternalism may in fact be worse than "hard" paternalism from a liberal perspective.

[1] Since even a total ban on anything only adds considerable costs (i.e. fines or criminal punishment) to procuring it, we can assume even traditionally paternalistic policies aren't 100% effective.

C - When Soft Paternalism is Harder than Hard Paternalism

Slovenian sociologist and philosopher Slavoj Žižek often employs the tale of "authoritarian parenting vs. permissive parenting" when criticising modern society[1]. For Žižek, authoritarian parenting simply represses the child's external behaviour; it doesn't expect the child to enjoy the limitations put on them, but to simply obey. So a child may be told that they have to come visit their grandma with their parents, whether they like it or not. Žižek contrasts this with "permissive" parenting, whereby the child is reminded of how much their grandmother loves them and how sad they would be if the child didn't come, to be finally left with the choice of staying at home or joining their parents. "Every child knows", says Žižek, "he is not being given any choice at all." Because in such a scenario, permissive parenting not only already knows what the child should choose, but forces the child to admit that choice internally. "You must visit your grandma, and you must be glad to do it!"

Žižek's example is of course exaggerated and provocative. There might be more than a dichotomy as a choice, or even worse for Žižek, the liberal parent might mean what they say: they might explain to the child that it would be nice of them to visit their grandma, but if the child ultimately says no, the liberal parent might actually accept that choice, hire a babysitter, and let them. Despite the exaggerated nature of the tale, it serves to illustrate a point: Despite libertarian paternalism's insistence that agents are influenced towards an option that they deem better *themselves* (and we will come to the question of how they claim to ensure this self-designated goal later), the insistence on Nudges as soft paternalism somewhat echoes the underhanded nature of permissive parenting. The idea that individuals not only fall in line with an option designated by an outside source, but then do so thinking it was their idea (i.e. that their preferences are *shaped* by that outside force, so that they would be different had that same force not been there) feels not like the triumph of freedom of choice and the victory of individual preferences Thaler and Sunstein herald, but its antithesis.

Thus the idea of Nudges *influencing* preferences, by affecting the internal decision-making mechanisms, feels more disruptive than traditional paternalism in the eyes of liberalism. While traditional liberalism cannot modify what the individual prefers (only block certain preferences from being achieved), libertarian paternalism changes the preferences themselves.

[1] Slavoj Žižek, *'You May!'*

It should be noted that libertarian paternalism's offense to liberalism in this case isn't due to its paternalistic nature, but due to the *nontransparent* changes to preferences that libertarianism accepts. The contradiction rises from the libertarian conception of choice as valuable regardless of the quality inherent to the choice. Liberalism, on the other hand, demands choices to be made while ensuring an equal moral worth and freedom *on fundamental rights*. The influence of nudging in this case can be likened to Bentham's panopticon: individuals aren't paternalistically forced to make a choice, but the idea that we could be watched at any moment (without knowing whether we are) changes our preferences in a nontransparent manner. Similarly, the way nudging can affect internal decision-making mechanisms, at least in certain situations[1], is deeply illiberal due to the nontransparent changes it brings to our behaviour.

In response to this criticism, Thaler and Sunstein underline the fact that their Nudges always leave the agents to make a bad choice if they truly wish to – the option to eat unhealthy food, smoke or spend unwisely is always there under nudge policies. The fact that an opt-out always exists thus seems to justify the influence Nudges otherwise wield – if the "true" preference of an agent is the choice the Nudge tries to steer people away from, they'll be able to choose it anyway[2]. In other words, there's always an escape hatch for an agent's individual preference if it is against the Nudge's direction.

While Sunstein and Thaler's defence seems plausible, it hides an inherent contradiction in the power it assumes Nudges have and the help "opt-outs" provide against these defaults. This problem is particularly apparent if one considers Nudges that instrumentalise one cognitive defect against another – in other words, Nudges form preferences in an equally irrational manner. If cognitive defects are so strong that individuals' "true" preference cannot shine through, using another cognitive defect to guide them in the choice architect's preferred direction doesn't help them reveal their true choice. Instead, it only guides them on the whims of a different cognitive defect. As law professor Gregory Mitchell puts it, "if a choice is as sticky as Sunstein and Thaler claim", where an individual's preference formation is too clouded by cognitive defects, "people who lack the means to determine their true preferences (…)

[1] This criticism applies only to specific nudges, as we will see in a moment.

[2] Thaler & Sunstein, *Nudge*, p. 5: "If people want to smoke cigarettes, to eat a lot of candy, to choose an unsuitable health care plan, or to fail to save for retirement, libertarian paternalists will not force them to do otherwise."

cannot make any real use of the opt-out provision (…) For these people, libertarian paternalism is just paternalism."[1]

The specific problem caused by opt-outs seems to point towards the problem not being with the concept of Nudges, but only specific ones. To be precise, the value liberalism places on individual preference clashes with libertarian paternalism when Nudges use cognitive biases. When Nudges only remove a cognitive defect but otherwise don't change preferences (such as a training), or when they create counter-incentives for better choices (which only changes the choice architecture on the outside – making a choice cognitively or economically costlier for example), Nudges don't seem to force an individual to like what they're doing – unlike permissive parenting. It's when one cognitive defect is wielded against another that the preferences Nudges claim to reveal become murky. In other words, Nudge theory seems to pose a problem for liberalism when it tries to manipulate us to make the "better" choice[2] (since we can assume that the individual would be unaware of a cognitive bias affecting them – they wouldn't make a faulty choice if they knew they were making it because of a cognitive bias) – if it only informs us or only counters a bias without animating another bias within us, it isn't powered by the same defects it aims to counteract. Thanks to this, it can be acceptable within the bounds of liberalism. Indeed, Mitchell keeps these "choice-independent" Nudges as he calls them exempt from his scathing criticism of libertarian paternalism, saying that choice-independent Nudges should be welcomed for promoting decision competence[3].

D - "As Judged by Themselves"?

A considerable portion of the liberal critique of libertarian paternalism hinges on the assumption that choice architects *impose* a preference to agents through Nudges. This is why Sunstein and Thaler have often responded to criticisms of Nudge theory that underlined the incompatibility between the idea that the choice architect should influence an agent's decision for the welfare of the agent and the principle of respecting individual preference by underlining, putting in italics, or making a chapter title out of a key part of their definition of libertarian

[1] Gregory Mitchell, *Libertarian Paternalism Is an Oxymoron*

[2] The lack of clarity on what "better" here means is a problem we will address immediately.

[3] Mitchell, *Libertarian Nudges*

paternalism: "influence choices in a way that will make choosers better off, *as judged by themselves*."

If the term that Sunstein and Thaler go back to again and again against their critics is indeed meaningful, our criticism of libertarian paternalism as a more insidious form of paternalism than its traditional kind (or a libertarianism that doesn't value liberalism's concerns) is suddenly deflated. If the choices towards which nudges guide agents are determined according to the informed preferences of the same agents, then we can't truly talk of Nudges imposing the will of the choice architect: the choice architect instead becomes a tool that the agents use to bind themselves to achieve the very goal they set themselves. Individual preferences remain at the forefront, liberalism prevails, and we may Nudge at will, even through defaults and (false) social norms.

If choices are truly determined by the welfare preferences of the agents, that is. While Sunstein and Thaler continuously underline that Nudges guide people to welfare goals they have set themselves, their writing on the subject explains shockingly little on *how* choice architects learn these goals. The examples they give are often on how to maximise goals that are assumed to be universal in our societies: eating more healthily, choosing an investment plan that will return more money, or reducing greenhouse gas emissions. Since the goals themselves are assumed to be universally shared, Sunstein and Thaler base the desirability of these goals on no survey or any preference expressed by someone other than themselves. While for most of their examples, the commonness of these goals given the circumstances might be true (especially if the goal itself can be expressed as a mean to an end like losing weight. Assuming those adopting the means already accept the same end, Nudges that guide people into choices that maximise their efficiency – like Nudges that speed up weight loss – intuitively make sense), it isn't necessarily true for every example[1]. And even if some of the cases for which Sunstein and Thaler suggest Nudges could somehow be known as universal *ex ante*, this tells us nothing on how choice architectures should set their goals for other cases where the agents' welfare

[1] Leaving aside any critique of the primacy of goals like more wealth being ideologically charged – since we had put that approach outside of our scope in the introduction by accepting a specific kind of rationality – we can still see that Sunstein and Thaler's suggestions to, for exampe, pollute less are not necessarily universally shared. A plastic company whose sole source of profit is the sale of single-use plastics certainly wouldn't prefer Nudges that help reduce personal waste.

preference is not known beforehand. Sunstein himself admits this problem, saying that for cases where Nudging would change a user's preference (i.e. cases where Nudges affect preference formation, thus forming an incompatibility with liberal principles), asking before instituting a Nudge isn't helpful and asking afterwards doesn't make sense (since preferences are already changed by the Nudge, so Nudges would be basing their legitimacy on a change in preferences they created themselves). After admitting this serious problem for the legitimacy of Nudges, Sunstein lists several potential solutions, without settling on one as a definitive answer. Thus, the problem of a Nudge's legitimacy for cases where the agents' preferences cannot be known *ex ante* remains unresolved[1].

In some cases, the *ex ante* preferences of agents may not even be legitimate even if we could know them. Sunstein and Thaler's *as judged by themselves* principle makes the assumption that the informed preferences of an individual on *any* subject is the absolute best option for them. Following the Elsterian model, sometimes collecting information on a decision is itself a bad course of action, if the optimal amount of information was less than what we try to collect. In other cases, the kind of rational decision-making may simply be useless for the choice in question: Elster gives the example of a career choice between becoming a lawyer or a forester: the life one will have in either scenario is too different to make any meaningful comparison. As a result, we're better off making a choice on secondary concerns (whether one can stay close to family as a forester, or if being a lawyer allows one to be closer to certain social groups…). As Elster puts it, "This way of deciding is as good as any", and enforcing informed preferences through nudging wouldn't be helpful in any way.

In even different cases, the vagueness of "increasing welfare" could bring back a question that came up when first defining paternalism: which increment of improving welfare is sufficient to justify being paternalistic? The problem is especially notable in cases where the agent's welfare contradicts the welfare of others to differing degrees: let us suppose that a certain household consumes considerably less electricity than the national average (say, 500 kWh per month when the national average is 860 kWh per month), and the government implements a nudge to reduce the national average through a disclosure nudge in their

[1] We will be discussing if democratic methods can be a viable solution, and if so under what conditions, in the following chapter.

electricity bills ("Our national average electricity consumption would need to be 700 kWh per month if we want to avoid climate catastrophe!"). The household, having seen that their consumption is far less than the needed number, could start consuming more, thus increasing their welfare (by feeling less obliged to consume less) while actually having the opposite effect that the government intended. In cases like these, the increment of change in welfare for the agent and the change to public welfare might make a clear-cut "better" choice much harder to find. All in all, we might be better off sticking to an instinctive choice rather than an informed preference in some situations.

A defender of libertarian paternalism might find this situation far from ideal, but nevertheless sufficient to legitimise their ideology. Sure, the scope of cases where Nudges that modify preferences can be used is more limited than Sunstein and Thaler's initial proposal, but it still exists, they might say. As long as choice architects stick to goals that we know *ex ante* as universally shared, we can still deem it legitimate to use Nudges that pit certain cognitive defects against another, even from a liberal perspective. This, they would say, still fits into the definition of Nudging for the benefit of an agent *as judged by themselves*. But even then, the idea seems dangerous due to the fallibility of choice architects. These choice architects are, after all, humans just like the agents they will be guiding through the Nudges they instate. Their decision-making process is no different than any other human, and is suspect to the same cognitive defects just as much as them. Therein lies the dilemma posed by treating cognitive defects as errors built into *every* human's decision making process: since they're errors that everyone shares, and since their effect may not be noticed at the time of taking a decision, the choice architect may make an error while designing the very Nudges that will correct the cognitive defects of others. Namely, they may think the Nudge they wish to instate that uses a cognitive defect against another allows achieving a universal welfare goal, but in truth the welfare they define is not universal. In such a case, they would be influencing the others' decision-making process with no legitimacy.

While such a scenario might seem too hypothetical or concern only fringe cases, real-life applications of Nudges have already run into this problem: in 2015, the UK's Behavioural Insights Team partnered with the Department for Work and Pensions as well as the Department of Health to Nudge people currently receiving the Employment and Social Allowance (ESA), a welfare payment for adults who cannot work either temporarily or permanently due to long-

term medical conditions or disabilities, to return to employment[1]. Their trials aimed to achieve this by having General Practitioners prescribe to ESA-recipients (i.e. people currently *unable* to work due to a medical condition or a disability) work coaches, or combining the health and employment information of ESA-recipients in a health and work passport. The idea seems to have been to influence ESA-recipients' motivation to seek a job, when the medical condition or disability that such a person has could make working directly harmful to them. Thus, it seems clear that the government's seemingly universal assumption of "everybody rationally wants to stop receiving benefits and work if possible" led to a Nudge that would directly reduce the welfare of some of the people Nudged. This example should illustrate the lack of a theoretical ground for basing nudges on welfare while remaining within a liberal framework. Perhaps this liberal framework should be used to build an alternative purpose to welfare.

E - Maximising Liberty and Not Welfare: A Liberal Alternative to Libertarian Paternalism

What would be the most premise-light principles to guide any Nudges we would need to make? While libertarian paternalism doesn't align with liberal principles, it has the advantage of offering a very tangible goal for Nudging: maximising an agent's welfare. Finding a similar principle from liberal principles proves to be more complicated than it initially seems, especially due to the complications in decision-making revealed by the Elsterian Rational Choice Model.

The knee-jerk response to "What is a liberal alternative to libertarian paternalism?" is quite simple: if libertarian paternalism aims to maximise welfare while respecting choice, liberal Nudges can set just maximising liberty as their principle. This proves to be a far murkier answer than libertarian paternalism's simple formula, however. What does liberty mean in this case, and what does maximising it entail?

One way of approaching liberty would be seeing it as maximising individual preference setting both quantitatively and qualitatively. Under such a definition, Nudges under liberal principles would aim to give as many choices as possible to individual, and make these choices as much meaningful as possible. However, a cursory look at how cognitive defects work under

[1] Matt Hancock (then-Minister for Cabinet Office), speech in *Future of Public Services Conference*

Elsterian Rational Choice Model shows the issue with such an approach: too many choices can actually run counter to meaningful choices (with too many options leading to analysis paralysis). Conversely, making choices meaningful might require reducing the number and possibilities of choices in some cases (precommitment requires an individual to be unable to turn from their decision to be meaningful, for example).

An alternative would be defining liberty as freedom from outside interference. Nudges in this case become synonymous with debiasing: they serve first and foremost to make sure our decision making processes are subject to no interference from outside influences that could skew it in one direction or another, and instead make sure that our actions remain in line with our desires and beliefs as much as possible. Such a design principle would be more than welcoming for Nudges that primarily neutralise and work around cognitive defects, and in fact would suggest implementing them as much as possible.

Such a liberal conception of Nudges would still fail to offer a justification for Nudges that use one cognitive defect against another, however. Since in such a Nudge it is still an external influence using one of your cognitive defects to influence your decision making process, these Nudges cannot be acceptable from a freedom-defending perspective. This seems to imply that in cases where some cognitive bias has to be exploited consciously, liberalism still can't justify any course of action. Gregory Mitchell, whose opinion on libertarian paternalism as strongly paternalistic we had referred to previously, seems to be of this opinion. He welcomes the use of a whole slew of "libertarian Nudges", but places them clearly within the confines of respecting freedom of choice. If one can accept acceptance of other influences in determining defaults, this seems like a plausible solution.

However, as stated in the beginning of this chapter, this dissertation's ideological boundaries are comprised of liberal *democracy*. While liberalism's premises are insufficient to justify deciding on a choice architecture in *every* way, principles of democracy might allow the use of Nudging as a viable policy tool in all its cognitive-defect-using glory. Furthermore, the implications of Nudging and the possibility of Nudging for democratic theory are equally worthy of consideration.

IV - Nudging for the People: Problems of Democracy

Liberalism and libertarian paternalism proved to be incompatible on several grounds: on one hand, libertarianism's lack of concern for the quality of choice lead to questionable nudges from a liberal perspective where the individual was influences in a deeply illiberal manner. Furthermore, the paternalistic aim of improving welfare, even when said welfare was defined by the agents, was hard to define and harder to implement without ideological errors. Thus, the concept of liberal nudging proved hard to be reconciled with liberalism and paternalism, much less the specific conception of libertarian paternalism Sunstein and Thaler adhere to. This chapter then asks the question of whether the democratic aim of public participation to government can offer some legitimacy to libertarian paternalism (for example, by showing a method to define welfare), or failing that, whether democratic aims can serve as an alternative design philosophy for nudging.

At first glance, the idea of the people correcting their own cognitive defects through self-imposed Nudges sounds plausible. If one were to interpret Nudging as an Odyssean self-binding where the people function analogously to a single person, the idea of democracy using the cognitive shortcomings of its citizens for better results (the ambiguousness of what that "better" means as discussed in the previous chapter notwithstanding) sounds like a novel and potentially very useful idea. However, the philosophy of exploiting our own cognitive defects in a democracy runs into two major problems: for Nudges that are compatible with libertarian paternalism (or either of its major components, libertarianism and paternalism), some of them raise major concerns for democratic principles. Other forms of nudging that can be based on democratic principles, however, run counter to the tenets of libertarian paternalism. These concerns, coupled with the liberal critique in the previous chapter, finally allow us a concrete answer on what cognitive defects and exploiting them in a liberal democracy can and should mean.

A - Libertarian Paternalistic Nudges: Incoherent with Democracy?

One way of seeing whether democratic principles can align with libertarian paternalist nudges is assuming libertarian and paternalist methods for nudging as a given, and seeing whether the result leads to acceptable outcomes from a democratic point of view. Doing so

reveals several incompatibilities between democracy and libertarianism, paternalism or their specific combination from the perspective of nudging.

1) Libertarianism Against Democracy

Libertarianism as a philosophy is far more permissive in terms of what private interests can do in a given situation. Libertarianism assumes that autonomy is assured when free market conditions and a maximum number of choices are secured, and nudging may be seen as a tool for such ends. However, choices with a multitude of options, when combined with certain nudges to specific directions, could be outright harmful for democracies.

What does the existence of Nudging imply for private interests? Above all, that they have certain methods to instrumentalise cognitive defects of their fellow citizens. Equipped with that knowledge, we can imagine individuals or smaller groups within a democracy to use Nudging for their own ends without trouble – a concerned group might launch a campaign to undo the biases of their fellow citizens about immigrants that are often hurt by false stereotypes, or (to reuse Sunstein and Thaler's very first example) a cafeteria owner might change the ordering on their menu to help their customers eat more healthily. If these examples are in line with public interest (as is the case with a debiasing campaign) or don't concern public interest (as is the case with a private cafeteria), Nudging by private groups is hardly a concern. This stops being true if one Nudges to further *their own ends* to the detriment of public interest, however.

Dark Patterns are emblematic of such situations. They're defined as a user experience crafted to trick the user into performing actions not in the user's own interest. The term originated with web UX (User eXperience) design, but has also been used to explain social capital and physical proximity.[1] The coiner of the term, UX Consultant Harry Brignull, gives several examples[2]: sneak into basket (where the website adds unexpected items to an online purchase unless opted out), roach motel (where getting into a situation is easier than getting out – i.e. the cognitive cost of going back is disproportionately higher), price comparison prevention (where websites make it harder to compare prices in order to stop users from making more

[1] Trice & Potts, 2018

[2] More examples can be found at https://www.darkpatterns.org/types-of-dark-pattern.

informed decisions) or hidden costs (such as websites listing one item but adding unexpected convenience fees right before payment). These examples are all notable for using one form of Nudging or another (sneak into basket is basically a default, a roach motel is the opposite of increasing ease, price comparison prevention practically aims to prevent debiasing, and hidden costs use limited attention spans), but unlike our examples so far, these Nudges benefit the interests of the website designer to the net loss of the user. The examples can clearly go beyond websites – and when they do, they're usually classified under unfair business practices. The advent of the Web has brought these practices into new areas, and has given them new forms, but the underlying principle (unconscionable circumstances to benefit a given party) remains the same. Our understanding of Nudging as policy design seems to only underline how these practices manipulate others. Luckily for most of these cases, similarity to unfair business practices has allowed several of these practices to already be prohibited. "Sneak into basket" methods are illegal in the EU following the Consumer Rights Directive, for example.

Dishonest business practices, though concerning, are hardly an objection to democracy by themselves, however. We might condemn these practices and refuse to do business with people that use manipulative Nudges for their benefit, just as we would condemn an abusive friend and distance ourselves from them. This by itself is a *deplorable* that doesn't necessarily run counter to democracy. It can, however, pose a serious threat to democracy if Nudges are used to promote private interests *to the detriment of public interest*, or to work around democratic decisions altogether. A notable example of such a situation once again concerns online activities, and the use of private data in particular. The General Data Protection Regulation (GDPR) is a Regulation adopted by the European Parliament (as well as the governments of all EU Member States through their approval in the Council of the European Union, also known as the Council of Ministers) in 2016 in response to growing concerns about the use of online personal data by third parties. The regulation specified that websites outside the European Union would now need the "informed consent" of all users before they could collect their data. Following the implementation of the regulation in 2018, all websites were required to ask users for their consent when placing cookies or collecting other data. Many websites changed their privacy policies, asking users to approve their data collection practices once again, under clearer terms, to ensure that the consent of these users was obtained.

Some websites, however, elected to go a different path. A considerable number of websites, instead of giving a clear choice between opting in to data processing or not, make one option far more complicated: opting in is a simple "I accept" while opting out is often hidden behind a secondary, unintuitive window (you must click on "Show purposes" then choose to opt out, which doesn't clearly evoke that the option to opt out will be there). In other cases, one must manually opt out of every single data vendor *individually* (and most websites work with hundreds of vendors), where a button labelled as "Reject all" doesn't actually reject all but only most of the vendors, and where reloading the page resets all preferences to opt-in for all vendors[1]. Similarly, some websites make the prospect of reaching a certain page more difficult. In some websites, when you land directly on a specific page (because you clicked on a specific news article, for example), opting out of data processing directs you automatically to the website's homepage, thus making access to the content you wanted to read harder unless you opt in to give away your personal data.

These dark patterns used by websites follow the letter of the GDPR: the user is offered a choice to give their consent or not, and the terms are explained in a detailed manner. But the spirit of the law is clearly violated by the Nudges towards accepting data processing. It is hard to call what is given "informed consent" when options are unequally presented, require dealing with deceiving wording and tedious design, or are simply manipulative. Most people (including myself) end up clicking on "I agree" out of frustration just to access the webpage, without considering the weight of the decision that is being made. The nefarious use of personal data for manipulating public opinion is already a well-known fact. As a result, any dark pattern that circumvents the GDPR is making it ever easier to undermine democracies and manipulate citizens, as it has been done so far by Cambridge Analytica.

The situation would be concerning even if dark patterns weren't used for this specific end: what is happening here, ties to data use for voter manipulation notwithstanding, is the *circumvention of democratically enacted laws*. In effect, dark patterns allow private interests to violate a law through an indirect manner: by making the implementation of a law fairly difficult if not impossible, these nefarious nudges allow private interests to trump democratic decisions.

[1] A demonstration of this dark pattern used by the VICE News website can be found at https://twitter.com/_gaeel_/status/1082702284746883072

Such a situation is unacceptable under any definition of democracy, as it gives some private interests an unequal preponderance. This is in direct opposition to the basic definition of democracy we had initially adopted, whereby some normative equality is assured between citizens. If Nudges can be used by certain private interests to trump not only other private interests but even public decisions, no normative equality can be assured.

2) Paternalism Against Democracy

If libertarianism, left unchecked with its singular concern for freedom of choice, corrupts nudges to nullify democracy, paternalism poses problems of a different sort. As we had noted in previous chapters, some early modern thinkers like Mill had valued political solutions to cognitive defects not only for their pragmatic benefits, but as educational tools as well. Following their reasoning, the form of government a people followed also had an effect on their character – i.e. their rational capabilities and in a sense, their "maturity". For these thinkers, representative government was preferable not only for its inherent virtues in allowing contributions from every individual, but also because people could learn to express themselves, debate their ideas, and be enlightened by the rational proposals that they would eventually encounter. In other words, democracy for these thinkers was valuable as an educational tool.

Nudging as a tool for paternalism effectively reduces these avenues for education. This is again valid mostly for nudges that use certain biases against others, and doesn't include every instance of nudging. But for those nudges that influence citizens' behaviour with no input from citizens, the opportunity for lessons that a public deliberation could offer is missed. In the case of a nudge that makes it harder to cancel gym subscriptions, for example, it is true that individuals will probably exercise more, thus public health will improve. But deciding to solve this public health issue through nudging instead of a discussion on *why* people don't commit to their gym memberships misses an opportunity to have a conversation with these citizens. Especially when such a conversation could reveal other structural issues (e.g. people not exercising due to stigma on exercising with an unfit body, or unforeseen difficulties in using the gym such as opening times), electing to nudge with a paternalistic instinct runs counter to the educational values of democracy. Even without a socially holistic view such as Mill's, one might find in this an erosion of democratic values.

3) Libertarian Paternalism Against Democracy

The previous arguments aimed to show the incoherencies between libertarianism (which doesn't care for the quality of choices made), paternalism (which can aim to influence to the detriment of meaningful growth) and democracy. One might nevertheless say that these dangers only show the need for *combining* these two philosophies to limit the vices of each other. This brings us to the question of whether libertarian paternalism, taken as a whole, can be compatible with democratic values.

How could libertarian paternalism be incompatible to democratic values? A major source of concern would be if nudging as a method runs counter to the intentions of democracy. Libertarian paternalism, by its definition, aims to lead individuals to better choices by intervening directly in their decision-making process. In other words, it aims to change their choice architecture, and in some forms, isn't interested in the input of the individuals in question. The problem arises, as with problems with liberalism, mostly due to nudges that pit certain cognitive biases against others. Nudges that merely inform or create competing incentives can be realistically approved in a democratic setting – the public may decide, after deliberation, to put reminders on cigarettes about their danger or make certain choices cognitively easier. However, deliberation on exploiting the public's own cognitive defects is a paradoxical task. If the public decides to "exploit" conformity to a social norm to nudge people to pay their taxes by saying that everyone else does, for example, it's clear that there is an inherent contradiction. Even discussing such an action would make the nonexistence of the desired social norm apparent, and thus make conformity impossible. A nudge on taxation that would make the cognitive weight of calculations easier, on the other hand (such as "taxing at the source" like in France) would be democratically feasible. Similarly, setting defaults for the public interest requires a democratic debate on what the default that benefits public interest is – and at that point, a rational discussion between the options is already had, and the status quo bias that would be used to Nudge people towards that option becomes unnecessary. Put another way, a default option decided democratically couldn't use the status quo bias, and an outright ban on other options or financial incentives toward the preferred option (both of which are solutions that are outside the scope of libertarian paternalism) would be more effective.

This of course assumes that Nudges would be decided democratically. If discussion to institute a Nudge isn't public, but is instead held behind closed doors by choice architects, then we see a deeper contradiction between Nudging and democracy: even if we assume that the choice architect has the best interests of the public at heart when deciding the Nudge, they would be deciding in a way that bypasses public debate. Nudging, in effect, becomes a method for working around competing social cleavages. While one may not use them with malicious intentions, the practice itself remains undemocratic.

Nudges that one cognitive defect against another seem impossible to implement in a democratic method. Either democratic decision-making nullifies their particular effectiveness, or a non-democratic deliberation is needed. It thus seems that in terms of implementation, these nudges seem incompatible with democracy. This doesn't say anything about nudging itself, but proves some incoherence between libertarian paternalism and democratic principles.

B - Democratic Nudges: Incoherent with Libertarian Paternalism?

If nudging is an interesting policy tool, but designing it around libertarian paternalism seems fruitless at best and detrimental to democracy at worst, basing nudges on democratic values could be an alternative. While it is true that Nudging is problematic from the perspective of liberalism because the direct link it assumes between informed preferences and welfare is uncertain at best, democratic principles do not pose such a contradiction. On the contrary, informed preferences align considerably better with democracy's desire to let all citizens have an equal say. If cognitive defects lead people to outcomes they wouldn't normally want, aren't policies designed for correcting them ultimately in service of revealing their informed preferences? Thus, the use of Nudges for *public* interest, (with various possible definitions as we will see shortly) could be a promising ground on which Nudges could be founded.

This seems to be an opinion reached by other scholars as well. Swiss law scholar Alexandre Flueckiger, in his article *Governing by "Nudge" – Using Cognitive Bias Instead of Legislation* makes a detailed analysis of Nudging and its design philosophy as defended by Sunstein and Thaler, and reaches a nuanced conclusion: while using Nudges *instead of* laws seems problematic (especially since it places Nudges outside the scope of rule of law), he ultimately concedes to the libertarian paternalistic argument that choice architecture must be designed – and that our choices are influenced one way or another. Individuals never truly

control their choices to a full extent, and some influence is inevitable and can be used for certain ends. However, Flueckiger refuses to agree that this means *the state* should take the decisions for the individuals. Rule of law and several other conditions ensure that this situation is avoided, and make Nudges permissible. The exact conditions Flueckiger gives are a long list, that we will not accept automatically as correct (as that would require justifying all of them separately), but as a demonstration of acceptable bases for nudging. These are: *a legal basis for Nudging*, *a public interest* that the Nudges aim to institute, *proportional application* of said Nudges, Nudging being done *in good faith*, and in a *transparent* and *non-discriminatory* manner[1]. The conditions Flueckiger sets can be seen as a democratic basis for Nudging, especially because public interest is put in the forefront[2]. However, using the term simply doesn't immediately prove that Nudges should aim to servce public interests, let alone that they can.

1) Public Interest: Hard to Define in a Democracy and Harder to Define for Nudges

What do we mean when we say that a Nudge should aim to fulfil a public interest? How can we decide on a public interest in a democratic framework, in a manner that allows a Nudge to be used for its fulfilment? The problem in answering this question is that a considerable body of work is dedicated to defining public interest, with mutually exclusive answers. A utilitarian conception would suggest the greatest utility for the greatest number as public interest (even for individuals who would come out worse off from it), while a Rousseauist conception of democracy might imagine public interest in a completely dissociated manner from private interests. An epistemic conception of democracy might place public interest in the proximity to the truth democratic processes can reach, while an egalitarian democrat might focus on social cohesion. In other interpretations, public interest might require an equality of resources even when this isn't in line with a private interest or individual welfare (famous example being Dworkin's study of Tiny Tim, where Tim's cheery disposition – i.e. higher welfare – does not justify not compensating his disability, and the paraplegic violinist where the community allocating funds to buy equipment to offset his disability is preferable even when a violin would

[1] Alexandre FLUECKIGER, *Gouverner par des « coups de pouce » (nudges) : instrumentaliser nos biais cognitifs au lieu de légiférer ?*

[2] It should however be noted that "public interest" as understood in law can be different from the philosophical conception of the term.

make him happier). It is impossible for all these conceptions to be right at the same time, and siding with one would demand a justification far out of the scope of this dissertation.

An alternative would be reaching a workable conception of public interest through the assumptions we hold about democracy. Such a conception would definitely be rougher than a more robust definition a specific conception of democracy would provide, but it would also prove to be far more compatible regardless of the specific conception of democracy we pick. With that in mind, we can turn to the basis we have defined for democracy in the beginning: that some *equal* right to participate is assumed between citizens. Such a definition demands neither that all citizens be materially equal, or even that every citizen's ideas be equally weighed. Under such a definition, even a Millian democracy where the "educated" opinions count more would be allowed, provided that an equality for each idea to *be present* is assured. Where does this basic assumption lead us regarding public interest, then?

If democracy is some sort of normative equality (in a sense of treatment as an equal and not equal treatment, following Dworkin's definition[1]) for all citizens, then the various competing public interests should all have a viable chance of being voiced. Thus, the variety of different conceptions seems to be a *feature* of democracies: while we can't define public interest with certainty, it is by allowing different conceptions to be voiced and to compete that we may get to a democratic definition of public interest. In other words, in a democracy, public interest could very well be *the result of competing cross-cutting cleavages in a society*[2]. Again, this tells us nothing about the essence of public interest: that is for the competing cleavages to decide through their competition. But this conception of democracy as a civic battleground for competing thoughts and interests is both intuitive and sufficiently functional for our purposes.

[1] Roland Dworkin, 1977, p. 370

[2] Robert A. Dahl, *A Preface to Democratic Theory*, p. 104-105: "The third defense, a popular one among American political scientists, draws upon the hypothesis of overlapping memberships. If most individuals in the society identify themselves with more than one group, then there is some positive probability that any majority contains individuals who identify themselves for certain purposes with the threatened minority. Members of the threatened minority who strongly prefer their alternative will make their feelings known to these members of the tentative majority who also, at some psychological level, identify themselves with the minority. Some of these sympathizers will shift their support away from the majority alternative and the majority will crumble."

Where do Nudges stand under such a conception of public interest? Can we now see what a Nudge for the public benefit may look like? Or does this conception of public interest pose a problem for Nudging?

2) Democratising Nudges: Resolving the Dilemma

The previous sections have argued against the use of libertarian paternalism as a framework for nudging. This seemed incoherent due to the specific combination of libertarian and paternalistic principles, which ran counter to liberal goals and worked against democratic processes. A disavowal of libertarian paternalism is not a disavowal of Nudging, however. The concept of Nudging i.e. designing policies *around* cognitive defects is not by itself abhorrent to liberal democracies. The problem seems to arise only when dealing with Nudges that *use* some cognitive defects against others. It is these Nudges in particular that pose a problem to liberalism – because as we argued in the end of the previous chapter, maximising liberty is an acceptable premise for Nudging instead of maximising welfare as revealed by informed preference. Once we remove the misguiding principles of libertarian paternalism, our aims (maximising meaningful choices) align with liberal principles. But this alignment is possible only if Nudges serve no further purpose than debiasing – thus, Nudges that aim to exploit certain cognitive defects to overcome others, since they do not improve meaningful choices, become unacceptable.

In a similar vein, limiting the scope of acceptable Nudges to exclude the exploitation of cognitive defects aligns considerably better with democratic principles. If libertarian paternalist Nudging seemed at odds with democratic forms of decision making, Nudging becomes far more acceptable to a democracy when we remove it. After all, the problems of paternalism only lie with the genuine intra-social conflicts Nudges could hide – and these hidden conflicts become a problem only when Nudges exploit cognitive biases, since these are the only ones that are inherently incompatible with democratic decision making. The same isn't true for Nudges that nullify cognitive defects or create incentives against them. A law that requires all campaigning parties to have the same opportunities in order to overcome any status quo biases, or policies that aim to increase the cognitive cost of buying cigarettes by making the process more difficult, or policies that decrease the cognitive cost of filing your taxes can all be approved through democratic decision making. The same is harder to say for policies that aim to use conformity

or lead to a certain default: conformity requires some ignorance on the subjects' end if it aims to *change* group behaviour (otherwise it is simply informing), and a default similarly gets its effectiveness from expecting that people *won't* bother to think about the default's subject. If a public deliberation is held, the usefulness of these Nudges becomes moot.

In fact, democracy seems to *give* Nudges a new legitimacy once we remove libertarian paternalism and its exploitative outlook. As Flueckiger points out, Nudges being done in service of a public interest is one of the key criteria that make them acceptable in a society under rule of law. *Democratising* Nudges, then, in the sense of using democratic methods to design and enact Nudges, is one way to solve our dilemma with them. After all, liberalism could find Nudges acceptable if the aim was changed to maximising liberty – but didn't give any particular reason *for* using them either. To return to liberal theory, it may find the persistence of cognitive defects much more bearable if the larger institutional design – the structure in which political decisions are made – is well designed to withstand any problems cognitive defects may bring. Thus, liberal theory seems to be *indifferent* at best to Nudging. Democracy, however, can turn that indifference to active use. If liberal theory only tolerates the use of Nudges as a not too illiberal policy tool, democracy can demand its use if it is in the public's interest to do so. It may become necessary to Nudge people and debias them if the democratic decision making process has found it to be necessary. This provides a much more enthusiastic basis for Nudging, while keeping that basis acceptable for liberal premises[1].

3) Nudging and Acknowledging the Weaknesses of Democracy

While democratic principles can serve as a fertile ground for designing nudges, the knowledge of dark patterns also raises a concern: Nudges can circumvent democratic politics altogether. While early modern conceptions of democracy accept that cognitive defects must be treated through education or incentives, there existed some belief that public decision, once taken, would stand inviolable. In other words, early modern conceptions of democracy could not predict these "soft" methods to circumvent laws for the benefit of private interests. The

[1] This basis is not enough to circumvent democratic decision-making however. The French government's decision to "nudge" citizens towards less carbon emissions with higher taxes on fuels, for example, was met with violent protests in 2018. Even nudges instituted in public interest need public approval.

paradigm proposed by cognitive biases and Nudging provides the ground for such strategies. As a result, democracy becomes forced to take into account cognitive defects in a new manner.

The previous section had shown a way for Nudges to democratised, while current discussion seems to show how democracies can be Nudged. The problem arises due to the fact that existing tools of government seem ill-equipped to handle these Nudges in their current form. In cases where a freedom of choice is assumed, dark patterns can effectively manipulate such a choice to the detriment of the agents, removing any authentic choice. In such a scenario, assuming that citizens consent normally is not only factually incorrect, but dangerous in terms of the ideas it promotes: such a situation effectively creates a double standard where the individual is assumed to be rational and actively giving consent, but the structure of the choice actually "forces" the individual to act in a certain way. The individual is in a state of superposition of assumed autonomy and factual manipulation. Such a state is clearly undesirable for a democracy for the incoherence it generates alone.

How, then, can we overcome this weakness of democracy revealed by dark patterns? The problems created by libertarian paternalism as a design philosophy could be overcome through a more democratic conception of Nudging. Similarly, a democracy that is designed with its cognitively vulnerable subjects can better handle dark patterns. Keeping cognitively vulnerable subjects in mind would mean that political institutions and their policy decisions would need be designed with potential workarounds through dark patterns in mind, as well as any ways in which political power could be abused by relying on irrational preferences spurred by cognitive biases. In the case of the GDPR, for example, a clear definition of informed consent and effective means to sanction service providers who manipulate users through dark patterns would be necessary for the regulation to work as intended. In fact, an outright ban of certain practices is probably needed. While this might seem like excessive regulation (especially for those of a libertarian bent like Sunstein and Thaler, who had proposed Nudging as a way to avoid regulations) to some, it has precedent in our existing democracies. Laws against libel and slander, for example, aim to protect the public persona of citizens against manipulation and propaganda. In these cases, slanderers are not simply allowed to distribute their manipulative opinion in the marketplace of ideas, but face discouraging sanctions for their behaviour. Similarly, a law against dark patterns or perverting the spirit of a law through cognitive defects may be necessary if democracy is to properly represent the will of the people.

Thus we see that our democracies not only require certain conditions to apply Nudges for public purposes, but require Nudges to maintain certain standards in order not to be corrupted.

Conclusion

Cognitive defects have been examined in a political context since at least the early modern period. The initial models understood these defects to be errors in judging the correct means to the correct ends, and the solutions devised by philosophers focused primarily on education, or creating new incentives. In other words, cognitive defects were entirely undesirable personal shortcomings that either needed to be corrected or managed.

The contemporary understanding of cognitive defects, informed by cognitive sciences, allows a different approach. This novelty is mostly due to cognitive defects being seen above as predictable, inherent deviations in our decision-making processes. Our perception of cognitive defects as measurable deviations common to every individual (if differing in intensity) allows a new, technical understanding. This understanding is technical because it turns cognitive defects into an instrument that can be used: Nudges build on the previous tools of education and incentives through debiasing trainings and cognitive incentives, but their primary innovation comes from the use of certain biases (mainly status quo bias and conformity bias) to offset other biases.

While this novelty has generated a lot of enthusiasm for Nudging and its primary design philosophy of libertarian paternalism, these new methods prove to be fairly controversial. The muddling between individual welfare and informed preferences, which Sunstein and Thaler try to equate, leads to a confused understanding of the ends Nudges serve. Liberal theory seems to suggest that the connection between informed preferences and welfare isn't as clear as they make it to be. Sometimes trying to find our informed preferences is directly detrimental to our welfare, and in other cases the freedom to create our own preferences is more important than any welfare goal a choice architect may deem fitting for us. A liberal aim of "maximising liberty" for Nudges makes more sense, but isn't too compelling for liberal theory.

Democratic values don't offer any valuable ground for libertarian paternalism either. On the contrary, the need for a public deliberation and a meaningful confrontation between ideas in a democracy makes the use of some Nudges for public interest impossible. Some

nudges (especially ones based on conformity) require some degree of ignorance from the recipient, thus a democratically-deliberated nudge would be paradoxical. As a result, it seems that neither liberal nor democratic values can condone the use of cognitive biases for political aims. Democratic concerns for the public interest do form a strong justification for debiasing attempts, however. This need for debiasing can form the basis for "democratising" Nudges, giving them a legitimate reason for existing in our societies.

While their use isn't desirable in a liberal democracy, the to use cognitive biases reveals some concerns for democracies: namely, that private interests can circumvent democratic decision making and even existing laws for their benefit and the detriment of everyone else. The new forms unfair practices can take in the marketplace of ideas must be accounted for if we want our democracies to remain healthy, and for the will of the people to be a genuine expression instead of the rubber-stamp approval of whoever manages to Nudge the sufficient number of people.

Future research can go in two separate directions. On the one hand, deepening our understanding of cognitive defects, and better defending ourselves against their use is an obvious direction to follow. More creative ways to overcome dark patterns (such as "gag rules" as defined by Stephen Holmes to avoid injuring "the cooperative spirit of organisation"[1]), and more interesting methods for democracies to Nudge themselves into better decisions (while still respecting the agents' freedom of choice) are left to be discovered. Most notably, the full scale of interactions within an Elsterian Rational Choice Model are not explored in the context of a democracy. Analysing other methods not detailed in this dissertation (self-binding on the scale of a society through hard constitutions, models of weakness of will in combatting commonly-accepted threats like climate change, situations of excess of will, etc.) can bring about a whole new array of Nudges that can allow us to face contemporary challenges, and help us understand better why we do certain things in a certain way.

Alternatively, a search for a whole new paradigm could be attempted. The sociological critique of cognitive sciences that we alluded to in the Introduction still merits some consideration. Such a critique would need to come up with an alternative model of decision-

[1] Stephen Holmes, Gag Rules or the Politics of Omission, in *Constitutionalism and Democracy*, p. 19-58.

making, that could perhaps attach a value to emotions without reducing them to a tool of rationality or a detriment to "correct" decisions. Works like Martha Nussbaum's *Upheavals of Thought: The Intelligence of Emotions*, that define emotions as "intelligent responses" with cognitive value, can find some merit for what we now deem to be irrational thinking. What we now deem to be a cognitive vulnerability may have its strengths, and if that is the case, finding these strengths is just as valuable as finding uses for cognitive biases.

Bibliography

ARIELY, Dan, *Predictably Irrational: the Hidden Forces that Shape Our Decisions*, New York: Harper Collins, 2008.

BENTHAM, Jeremy, *A Comment on the Commentaries and A Fragment on Government*, ed. J. H. Burns and H. L. A. Hart, Oxford: Clarendon Press, 1977.

BENTHAM, Jeremy, *An Introduction to the Principles of Morals and Legislation*, ed. J. H. Burns and H. L. A. Hart, Oxford: Clarendon Press, 1970.

BENTHAM, Jeremy, *Catalogue of the Manuscripts of Jeremy Bentham in the Library at University College, London.* Compiled by A. Taylor Milne, 1937.

BENTHAM, Jeremy, STARK, W. (ed.), *Jeremy Bentham's Economic Writings*, volume III., London: George Allen & Unwin, 1954.

BENTHAM, Jeremy, *Of the Limits of the Penal Branch of Jurisprudence*, ed. P. Schofield, Oxford: Clarendon Press, 2010.

BENTHAM, Jeremy, *Theory of Legislation, by Jeremy Bentham. Translated from the French of Etienne Dumont, by Richard Hildreth*, 2 vols., rept. Bristol: Thoemmes Continuum, 2004.

CAMERER, Colin, ISSACHAROFF Samuel, LOEWENSTEIN George et al. "Regulation for Conservatives: Behavioral Economics and the Case for "Asymmetric Paternalism", *University of Pennsylvania Law Review*, 2003, vol 151, p. 1211–1254.

DAHL, Robert A., *A Preface to Democratic Theory*, 50th Anniversary Edition, London: the University of Chicago Press, 2006.

DEMARQUE Christophe, CHARALAMBIDES, Laetitia, HILTON, Denis J., et al. "Nudging sustainable consumption: The use of descriptive norms to promote a minority behavior in a realistic online shopping environment", *Journal of Environmental Psychology*, 2015, vol. 43, p. 166-174.

DWORKIN, Ronald, *Sovereign Virtue: the Theory and Practice of Equality*, Cambridge, Massachusetts: Harvard University Press, 2002

DWORKIN, Ronald, *Taking Rights Seriously*, Cambridge: Harvard University Press, 1977.

ELSTER, Jon, *Reason and Rationality*, Steven Rendall (translated by), New Jersey: Princeton University Press, 2008.

ELSTER, Jon, *Solomonic Judgements: Studies in the Limitations on Rationality*, Paris: Maison des Sciences de l'Homme, 1989.

ELSTER, Jon, *Ulysses Unbound: Studies in Rationality, Precommitment and Constraints*, Cambridge: Cambridge University Press, 2000.

FLUECKIGER, Alexandre. "Gouverner par des « coups de pouce » (nudges) : instrumentaliser nos biais cognitifs au lieu de légiférer ?" *Les Cahiers de droit*, 2018, vol. 59, n° 1, p. 199-227

FOUCAULT, Michel, "The History of Sexuality," in *Power/Knowledge. Selected Interviews and Other Writings 1972-1977*, edited by Colin Gordon, Pantheon Books New York 1980.

GAVIL, Andrew I., FIRST, Harry, *The Microsoft Antitrust Cases: Competition Policy for the Twenty-first Century*. Cambridge, Massachusetts, 9 December 2014.

HANCOCK, Matt, "Future of Public Services Conference: Matt Hancock Speech", *Local Government Association Conference*, 21 October 2015, Speech by the Minister for Cabinet Office. Retrieved from https://www.gov.uk/government/speeches/future-of-public-services-conference-2015-matt-hancock-speech.

HOLMES, Stephen. "Gag Rules or the Politics of Omission." Chapter. In *Constitutionalism and Democracy*, edited by Jon Elster and Rune Slagstad, Cambridge: Cambridge University Press, 1988, p. 19–58.

KAHNEMAN, Daniel, TVERSKY, "Subjective probability: A judgment of representativeness", *Cognitive Psychology*, 1972, vol. 3, n° 3, p. 430-454.

KIRSCH, Irving, "Response expectancy as a determinant of experience and behavior". *American Psychologist*, 1985, vol. 40 (11), p. 1189–1202.

KYMLICKA, Will, *Contemporary Political Philosophy: An Introduction*, 2nd Edition, Oxford: Oxford University Press, 2002.

LOCKE, John, *An Essay Concerning Human Understanding*, Peter H. Nidditch (ed.), Oxford: Clarendon Press, 1975 (The Clarendon Edition of the Works of John Locke).

LOCKE, John, *"Some Thoughts Concerning Education" and "The Conduct of the Understanding"*, Ruth W. Grant and Nathan Tarcov (eds), Indianapolis: Hackett Publishing Co., 1996.

MANDLER, George, "Origins of the cognitive (r)evolution", *Journal of the History of the Behavioral Sciences*, 2002, vol. 38, n° 4, p. 339-353.

MEHTA, Uday Singh, *Liberalism and Empire: A Study in Nineteenth-Century British Liberal Thought*, Chicago: The University of Chicago Press, 1999.

MILL, John Stuart, *The Collected Works of John Stuart Mill*, John M. Robson (ed.), London: Routledge and Kegan Paul.

MITCHELL, Gregory, "Libertarian Nudges", *Virginia Public Law and Legal Theory Research Paper*, November 1, 2017, No. 2017-65.

MITCHELL, Gregory, "Libertarian Paternalism is an Oxymoron", *Northwestern University Law Review*, Vol. 99, No. 3, 2005.

MOREWEDGE, Carey K., YOON Haewon, SOPELLITI Irene, et al. "Debiasing Decisions: Improved Decision Making With a Single Training Intervention." *Policy Insights from the Behavioral and Brain Sciences*, October 2015, vol. 2, n° 1, p. 129–40.

MORGAN, Glyn. "The Mode and Limits of John Stuart Mill's Toleration." *Nomos*, vol. 48, 2008, p. 139–167.

NUSSBAUM, Martha, *Upheavals of Thought: The Intelligence of Emotions*, New York: Cambridge University Press, 2001.

ROBBINS, Lionel, *An Essay on the Nature & Significance of Economic Science*, London: Macmillan, 1932.

StatCounter Global Stats. *Browser Market Share Europe | StatCounter Global Stats*. [online] n.d. Available at: http://gs.statcounter.com/os-market-share/all/europe/2010 [Accessed 20 May 2019].

StatCounter Global Stats. *Operating System Market Share Europe | StatCounter Global Stats*. [online] n.d. Available at: http://gs.statcounter.com/os-market-share/all/europe/2010 [Accessed 20 May 2019].

SUNSTEIN, Cass, THALER, Richard, "Libertarian Paternalism is Not an Oxymoron", *University of Chicago Law Review*, 2003, vol. 70, n° 4, p. 1159–1202.

SUNSTEIN, Cass, "Nudging: A Very Short Guide", *Journal of Consumer Policy*, November 2014, vol. 37, n° 4, p. 583-588.

SUNSTEIN, Cass, "The Ethics of Nudging", *Yale Journal on Regulation*, 2015, vol. 32, n°2.

THALER, Richard H., SUNSTEIN, Cass, *Nudge: Improving Decisions About Health, Wealth and Happiness*, Michigan: Caravan, 2008.

THALER, Richard, SUNSTEIN, Cass, "Libertarian Paternalism", *The American Economic Review*, 2003, vol. 93, p. 175–79.

TRICE, Michael, POTTS, Liza, "Building Dark Patterns into Platforms: How GamerGate Perturbed Twitter's User Experience", *Present Tense: A Journal of Rhetoric in Society*, January 21, 2018, vol. 6, n° 3.

WILK, James, "Mind, nature and the emerging science of change: An introduction to metamorphology.", in G. Cornelis; S. Smets; J. Van Bendegem (eds.), *EINSTEIN MEETS MAGRITTE: An Interdisciplinary Reflection on Science, Nature, Art, Human Action and Society: Metadebates on science*, 6, Springer Netherlands, 1999, pp. 71–87.

WOLFF, Johnathan, *Ethics and Public Policy: A Philosophical Inquiry*, New York: Routledge, 2011.

ZIZEK, Slavoj, "You May!". *London Review of Books*, 1999 vol. 21 n° 6, p. 3-6.